

On the Frequency of Words Used in Answers to Explain in Plain English Questions by Novice Programmers

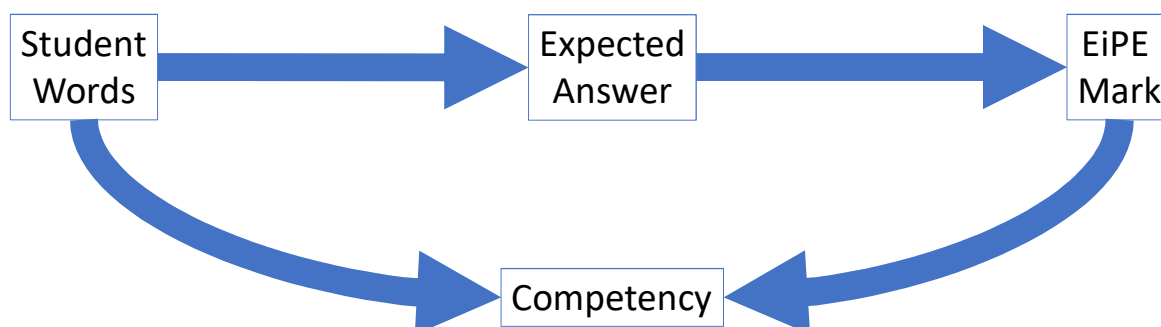
Authors: Thomas Pelchen (Presenter/PhD Candidate) &
Raymond Lister (Supervisor)

2019 ACE Conference, 29 January – Macquarie University

Hello and welcome to this presentation on our paper **[READ]**

My name is Thomas Pelchen, I am the presenter and one of the authors of this paper.
Raymond lister, my supervisor and the other author on this paper unfortunately couldn't
make it here today.

EiPE Questions and their Words



- The relationship between the marks from Explain in Plain English (EiPE) questions and a student's competency has been well established.

The best way to explain reasoning for this study is to just tell you how we came up with the idea for this study in the first place.

To set the scene, picture two researchers in a room full of marked exam papers, looking at spreadsheet of marks.

Whilst looking at the grading scheme for the Explain in Plain English questions, an idea presented itself. The process of marking an Explain in Plain English question is that we first **[ANIM]** look at the words in an answer, then we compare **[ANIM]** it to the expected answer to give it a mark **[ANIM]** depending on whether it matched what we were expecting.

Now, **[ANIM]** the relationship between the marks of Explain in Plain English questions and a student's competency is well established. **[ANIM]** So we can say that we use the marks to determine the competency of the student.

So what if we flipped this on its head? **[ANIM]** Could we use the students words to determine the competency?

The Research Question

- Among students who answer a specific EiPE question correctly, is there a difference between the answers of students who did well on other EiPE questions and students who did not do as well on other EiPE questions?

To this end, we proposed the following research question **[READ]**

Correct Relational Answers Only

- For this question, we are interested in correct answers only.
 - For an exam question to be marked correct, the answer needed to be 'relational' as per the SOLO taxonomy.
- We are interested in the words students use when answering those questions.
- Whether the total mark they received for the EiPE section can be identified through the words that they used.

[READ]

As there is only one way to generally give a correct answer, and many different ways to give a wrong one, it would be interesting to see if there is any variance in the correct answers.

Institutional Context

- Number of Students: 334 students*
- Number of EiPE Exam Questions: 12
- Number of Answers: 3636

*31 students scored zero on the EiPE section and are not included.

So, to give you some background on the cohort. The answers used in this study come from the final written exam of an introductory programming (CS1) class at the University of Technology Sydney.

It was not necessary for these students to take the final exam to pass the subject, 334 students undertook the final exam **[READ *]**.

In order to get the data for the answers, we typed up the answers to these 12 questions: approximately 3600 answers.

Bands?

- We categorised the students into 3 main bands:

Band 12	High - 36 students who answered all EiPE questions correctly
Band 9	Medium - the 39 students who answered 9 out of the 12 EiPE questions correctly
Band 4-7	Low - The 69 students who answered 4 to 7 of the 12 EiPE questions correctly

We split the cohort into three main bands in response to the research question. One that targets the students that did well, the 12 band; one that targets the students that not did so well, the 4-7 band; and one right between them, the 9 band.

[For justification if asked] In order to conduct a reasonable test, we needed at least 20 students in each band. So we expanded the lowest band to include at least 20 students with the harder questions. The 9 band was chosen for both its population size and its distance from the 12 and 4-7 bands.

The Tests

	No. of Students who used word W in a correct relational answer	No. of Students who did NOT use word W in a correct relational answer
Band x		
Band y		

- A Chi-squared test was used on the above table to determine its significance.

So this is what we did for the tests. For each word, lets call it 'W', we plotted the following **READ**

We then ran a Chi-squared test on that table to determine if the word was statistically significant.

[If asked]

Average words: We ran ANOVA tests on the average length of each answer, apart from one question, the results were not significant.

Corrections: if needed, we ran a Fisher Exact test as a correction for 2x2 tables and a Fisher-Freeman-Halton exact test for 3x2 tables.

Q34 “find”

```
public static int q34(int data[], int x )
{
    int z = -1;

    for (int i=0; i < data.length; i++ )
    {
        if ( data[i] == x )
            z = i;
    }

    return z;
}
```

36 Students who got 12				37 Students who got 9				35 Students who got 4-7	
Word	% of use			Word	% of use			Word	% of use
THE	97			THE	89			THE	89
RETURN	94			RETURN	81			RETURN	66
ARRAY	92			ARRAY	84			ARRAY	80
POSITION	86			POSITION	54			POSITION	69
'X'	86			'X'	76			'X'	86
-1	83			-1	76			-1	51
IF	83			IF	70			IF	69
IN	83			IN	76			IN	60
'DATA'	72			'DATA'	59			'DATA'	46
VALUE	64			VALUE	68			VALUE	74
OF	64			OF	46			OF	60
FOUND	61			FOUND	41			FOUND	11
NOT	53			NOT	46			NOT	26
SEARCH	25			SEARCH	22			SEARCH	6
ELEMENT	25			ELEMENT	8			ELEMENT	26
LAST	25			LAST	11			LAST	6
FOR	25			FOR	30			FOR	9
OR	25			OR	27			OR	11
EQUAL	19			EQUAL	16			EQUAL	37
OCCUR	17			OCCUR	3			OCCUR	6
INTEGER	17			INTEGER	14			INTEGER	9
OTHERWISE	17			OTHERWISE	24			OTHERWISE	14
THAT	14			THAT	19			THAT	23
CODE	14			CODE	8			CODE	9
WHERE	14			WHERE	8			WHERE	3
MATCH	14			MATCH	3			MATCH	20
NO	11			WHICH	3			NO	9

Now the next few slides will detail the results of our study. To help you make sense of them, I will now explain them.

On the left half **[ANIM]** we have the question used in the exam on the top and underneath is where I will be putting the possible answer.

On the right half **[ANIM]** we have the list of correct words. Now this is separated into three main groups **[POINT]** the left is the 12 band, the right is the 4-7 band and the middle is the 9 band.

In these lists the number next to the word indicates the percentage of the bands students who used that word in their answers. **[ANIM]** Have a look at the top word 'The', the table is showing that, of the students in band 12, 97% of them used the word 'the' in their answer.

The words in the 12 band are sorted in descending order by their percentage of use. For ease of reading and comparison, the words of the other bands are listed to match the 12 band.

Q34 “find”

```
public static int q34(int data[], int x )
{
    int z = -1;

    for (int i=0; i < data.length; i++ )
    {
        if ( data[i] == x )
            z = i;
    }

    return z;
}
```

IT

36 Students who got 12		37 Students who got 9		35 Students who got 4-7	
Word	% of use	Word	% of use	Word	% of use
THE	97	THE	89	THE	89
RETURNS	94	RETURN	81	RETURN	66
ARRAY	92	ARRAY	84	ARRAY	80
POSITION	86	POSITION	54	POSITION	69
'X'	86	'X'	76	'X'	86
-1	83	-1	76	-1	51
IF	83	IF	70	IF	69
IN	83	IN	76	IN	60
'DATA'	72	'DATA'	59	'DATA'	46
VALUE	64	VALUE	68	VALUE	74
OF	64	OF	46	OF	60
FOUND	61	FOUND	41	FOUND	11
NOT	53	NOT	46	NOT	26
SEARCH	25	SEARCH	22	SEARCH	6
ELEMENT	25	ELEMENT	8	ELEMENT	26
LAST	25	LAST	11	LAST	6
FOR	25	FOR	30	FOR	9
OR	25	OR	27	OR	11
EQUAL	19	EQUAL	16	EQUAL	37
OCCUR	17	OCCUR	3	OCCUR	6
INTEGER	17	INTEGER	14	INTEGER	9
OTHERWISE	17	OTHERWISE	24	OTHERWISE	14
THAT	14	THAT	19	THAT	23
CODE	14	CODE	8	CODE	9
WHERE	14	WHERE	8	WHERE	3
MATCH	14	MATCH	3	MATCH	20
NO	11	WHICH	3	NO	9

With each question we look at today, I will give an example answer for it.

A good answer to this question would that it “It finds the value in the array”. Some of you may have spotted this already, but you can actually see what the most common answers is to this question just by looking at the percentage of the words used. .”

[ANIM] This question for example: “It returns the position of ‘x’ in the array

One other thing, note the quotation marks around the variable ‘data’ in the 12 table **[POINT]**. Words in quotation marks indicate the name of a variable, this is to help the reader identify what is a variable name and what is a word.

Q34 “find”

```
public static int q34(int data[], int x )
{
    int z = -1;

    for (int i=0; i < data.length; i++ )
    {
        if ( data[i] == x )
            z = i;
    }

    return z;
}
```

To be marked correct, a student did **NOT** have to specify that:

1. If the search value is found it returns its position
2. The method returns -1 if not found
3. If it finds more than one value it returns the position of the final one

36 Students who got 12		37 Students who got 9		35 Students who got 4-7	
Word	% of use	Word	% of use	Word	% of use
THE	97	THE	89	THE	89
RETURN	94	RETURN	81	RETURN	66
ARRAY	92	ARRAY	84	ARRAY	80
POSITION	86	POSITION	54	POSITION	69
'X'	86	'X'	76	'X'	86
-1	83	-1	76	-1	51
IF	83	IF	70	IF	69
IN	83	IN	76	IN	60
'DATA'	72	'DATA'	59	'DATA'	46
VALUE	64	VALUE	68	VALUE	74
OF	64	OF	46	OF	60
FOUND	61	FOUND	41	FOUND	11
NOT	53	NOT	46	NOT	26
SEARCH	25	SEARCH	22	SEARCH	6
ELEMENT	25	ELEMENT	8	ELEMENT	26
LAST	25	LAST	11	LAST	6
FOR	25	FOR	30	FOR	9
OR	25	OR	27	OR	11
EQUAL	19	EQUAL	16	EQUAL	37
OCCUR	17	OCCUR	3	OCCUR	6
INTEGER	17	INTEGER	14	INTEGER	9
OTHERWISE	17	OTHERWISE	24	OTHERWISE	14
THAT	14	THAT	19	THAT	23
CODE	14	CODE	8	CODE	9
WHERE	14	WHERE	8	WHERE	3
MATCH	14	MATCH	3	MATCH	20
NO	11	WHICH	3	NO	9

[READ], [ANIM], [READ], [ANIM], [READ], [ANIM], [READ],

Now it is these three points that show significance, as this is where the words that are used really show the difference between the bands.

Note that action of ‘finding’ is mentioned in all three points. **[CLICK]**

Q34 “find”

```
public static int q34(int data[], int x )
{
    int z = -1;

    for (int i=0; i < data.length; i++ )
    {
        if ( data[i] == x )
            z = i;
    }

    return z;
}
```

To be marked correct, a student did **NOT** have to specify that:

1. If the search value is **found** it returns its position
2. The method returns -1 if not **found**
3. If it **finds** more than one value it returns the position of the final one

36 Students who got 12		37 Students who got 9		35 Students who got 4-7	
Word	% of use	Word	% of use	Word	% of use
THE	97	THE	89	THE	89
RETURN	94	RETURN	81	RETURN	66
ARRAY	92	ARRAY	84	ARRAY	80
POSITION	86	POSITION	54	POSITION	69
'X'	86	'X'	76	'X'	86
-1	83	-1	76	-1	51
IF	83	IF	70	IF	69
IN	83	IN	76	IN	60
FOUND	61%	FOUND	41%	FOUND	11%
NOT	53	NOT	46	NOT	26
SEARCH	25	SEARCH	22	SEARCH	6
ELEMENT	25	ELEMENT	8	ELEMENT	26
LAST	25	LAST	11	LAST	6
FOR	25	FOR	30	FOR	9
OR	25	OR	27	OR	11
EQUAL	19	EQUAL	16	EQUAL	37
OCCUR	17	OCCUR	3	OCCUR	6
INTEGER	17	INTEGER	14	INTEGER	9
OTHERWISE	17	OTHERWISE	24	OTHERWISE	14
THAT	14	THAT	19	THAT	23
CODE	14	CODE	8	CODE	9
WHERE	14	WHERE	8	WHERE	3
MATCH	14	MATCH	3	MATCH	20
NO	11	WHICH	3	NO	9

$p = 0.001$ $p = 0.005$

As the code makes no mention of the word ‘found’, the use the word ‘found’ in an answer is an abstraction from the question. When looking at its usage across the bands [ANIM], note the percentages.

[POINT] See the difference between the words usage in the 4-7 band? 11% compared to 41 and 61 percent.

This relationship between the 4-7 and 9 bands [ANIM] is significant at $p = 0.005$ and the 4-7 and 12 bands [ANIM] with $p = < 0.001$.

Q34 “find”

```
public static int q34(int data[], int x )
{
    int z = -1;

    for (int i=0; i < data.length; i++ )
    {
        if ( data[i] == x )
            z = i;
    }

    return z;
}
```

To be marked correct, a student did **NOT** have to specify that:

1. If the search value is **found** it returns its position
2. The method returns -1 if not **found**
3. If it **finds** more than one value it returns the position of the final one

36 Students who got 12		37 Students who got 9		35 Students who got 4-7	
Word	% of use	Word	% of use	Word	% of use
THE	97	THE	89	THE	89
RETURN	94	RETURN	81	RETURN	66
ARRAY	92	ARRAY	84	ARRAY	80
POSITION	86	POSITION	54	POSITION	69
'X'	86	'X'	76	'X'	86
-1	83	-1	76	-1	51
IF	83	IF	70	IF	69
IN	83	IN	76	IN	60
'DATA'	72	'DATA'	59	'DATA'	46
VALUE	64	VALUE	68	VALUE	74
OF	64	OF	46	OF	60
FOUND	61	FOUND	41	FOUND	11
NOT	53	NOT	46	NOT	26
SEARCH	25	SEARCH	22	SEARCH	6
ELEMENT	25	ELEMENT	8	ELEMENT	26
LAST	25	LAST	11	LAST	6
EQUAL 19%		EQUAL 16%		EQUAL 37%	
INTEGER	17	INTEGER	14	INTEGER	9
OTHERWISE	17	OTHERWISE	24	OTHERWISE	14
THAT	14	THAT	19	THAT	23
CODE	14	CODE	8	CODE	9
WHERE	14	WHERE	8	WHERE	3
MATCH	14	MATCH	3	MATCH	20
NO	11	WHICH	3	NO	9

p = 0.044

Conversely, we see that the 4-7 band uses words more like ‘equal’ in their answers [ANIM]. A glance shows you that ‘equal’ is used in [POINT] 37% of the answers by the 4-7 band compared with 16 and 19 percent for the 9 and 12 bands respectively.

In terms of significance [ANIM], the relationship between 9 and 4-7 is significant as $p = 0.044$. But not at all between the 9 and 12 bands.

This adds towards a finding throughout the questions. That the higher bands are more focused on the result itself, whilst the lower bands are concerned with the process of reaching that answer. Essentially, the lower students are unable to decouple their working out from the answer.

To investigate, we went back to look at the students answers. We were indeed able to confirm that the higher performing students talk about the value being ‘found’, compared to the lower performing students who talk about the value in the array being ‘equal’ to the search value.

Q28 “large3”

```
if ( a < b)
{
    if ( b < c)
        System.out.println (c);
    else
        System.out.println (b);
}
else
{
    if ( a < c)
        System.out.println (c);
    else
        System.out.println (a);
}
```

D?

It prints the largest variable.

12 – It prints the largest value in variables ‘a’, ‘b’ and ‘c’.

9/4-7 – It prints the largest value or variable

36 Students who got 12		38 Students who got 9		50 Students who got 4-7	
Word	% of use	Word	% of use	Word	% of use
THE	97	THE	100	THE	96
PRINT	92	PRINT	95	PRINT	88
‘A’, ‘B’, ‘C’		p = 0.003		p = < 0.001	
81%		47%		‘A’, ‘B’, ‘C’	
				44%	
VALUE	72	VALUE	61	VALUE	72
VARIABLE	67	VARIABLE	39	VARIABLE	48
AND	58	AND	24	AND	38
OF	56	OF	39	OF	30
OUT	36	OUT	45	OUT	46
IT	31	IT	24	IT	36
IN	28	IN	8	IN	24
HIGHEST	17	HIGHEST	16	HIGHEST	16
STORE	17	STORE	3	STORE	4
THREE	14	THREE	16	THREE	14
OR	14	OR	8	OR	4
WITH	11	WITH	11	WITH	20
INTEGER	11	INTEGER	3	CODE	20
CODE	11	CODE	8	WILL	12
WILL	8	3	3	3	8
3	8	NUMBER	13	NUMBER	14
NUMBER	8	THIS	3	THIS	4
HAS	8	BETWEEN	8	BETWEEN	2
THIS	8	BIGGEST	11	THAT	2
BETWEEN	8	‘D’	5	BIGGEST	14
THAT	8	WHICH	3	WHICH	2
BIGGEST	8	RETURN	3	RETURN	2
‘D’	8	CONTAIN	3	FIND	8
WHICH	6	BY	3	ELEMENT	2
FOUR	6	MAXIMUM	3	COMPARE	2
RETURN	6	LETTER	3	MAXIMUM	2
CONTAIN	6	FROM	3	LETTER	6
FIND	6	AMONG	3	FROM	2
BY	3	TO	11	AMONG	2

A correct answer to this question would be that it **[ANIM]** **[READ]**

What seems to be a trend in this and many other questions we looked at, is that the higher bands tend to directly reference the variable. **[ANIM]** See that the 12 band refers to the variables ‘a’, ‘b’ and ‘c’ in 81% of their answers while 9 and 4-7 refer to it in 47 and 44 percent respectively.

These values are significant **[ANIM]** with $p = 0.003$ between 12 and 9 and **[ANIM]** $p < 0.001$ between 12 and 4-7.

Although we have no definite conclusion as to why this is the case for most of the questions, as the reasoning may change depending on the question, we can make an observation about this one.

It is entirely possible that the 12 band prefer giving context in their answer **[ANIM]**, “it prints the largest value in variables ‘a’, ‘b’ and ‘c’”. And, outside of perhaps **[ANIM]** “it prints the largest value or variable”, we suspect that 4-7 and 9 bands use a variety of different ways to answer this question, ways which are not immediately obvious in this table.

On a side note, one of my absolute favourite things about the answers to this particular

question is the use of the variable **[ANIM]** 'D' in their answer. Something that still has us scratching our heads.

Q32 “sum”

```
int z = 0;

for (int i=0 ; i<x.length ; ++i )
    z = z + x[i];

System.out.println(z);
```

It prints the sum of all the values in the array.

36 Students who got 12			38 Students who got 9			36 Students who got 4-7	
Word	% of use		Word	% of use		Word	% of use
THE	97		THE	100		THE	92
SUM	89		SUM	84		SUM	64
PRINT	86		PRINT	63		PRINT	72
OF	86		OF	87		OF	83
ARRAY	86		ARRAY	92		ARRAY	92
'X'			'X'			'X'	
69%			47%			47%	
ELEMENT	33		ELEMENT	13		ELEMENT	17
IT	28		IT	24		IT	42
AND	25		AND	11		AND	22
'Z'			p = 0.023			p = 0.003	
19%			13%			44%	
UP	14		UP	8		UP	6
TOTAL	14		TOTAL	11		TOTAL	14
ADDED	14		ADDED	16		ADDED	22
FIND	11		FIND	3		IS	8
IS	11		IS	3		TO	19

For this question, a correct answer to this question would be **[ANIM]** **[READ]**

Inspecting this table shows a interesting result, lets have a closer look at the two main variables 'X' and 'Z' **[ANIM]**

Although not statistically significant **[ANIM]**, we do see a 22% increase of the 'X' variables use in the answers of the 12 band, when compared to the 4-7 and 9 bands **[ANIM]**.

However, it is when we get to the variable 'Z' where we get to that interesting result **[ANIM]**. Band 4-7 shows a more than double the use of the auxiliary variable 'Z' when compared to the 9 and 12 bands. These relationships are significant, where between **[ANIM]** the 4-7 and 9 bands $p = 0.003$ and **[ANIM]** between the 4-7 and 12 bands $p = 0.023$.

This again, brings me back to the earlier point where the lower bands are more focused on the calculation of a result (which 'Z' is used for), while the higher bands are more focused on the result itself.

Implications of Results

- Students who do well are more likely to use words that abstract beyond the code.
- That in the words used when answering EiPE questions, the student's level of development can be identified.
- That you can get more out of EiPE questions than whether or not they have gotten the question correct.
- We are not claiming that this could be used as an assessment tool, but rather an aid for educators to identify struggling students.
 - As the usefulness of these findings are limited by the flexibility allowed for a correct response to the question.

To answer our earlier proposed research question, we did indeed find that when answering an EiPE question, there is a difference in the answers between students who performed well and students that did not perform so well.

Particularly that **[ANIM]**, **[READ]**

Which can be logically extended and leads us to our second point **[ANIM]**, **[READ]**

And **[ANIM]**, **[READ]**

Now, I wish to stress the last point that **[ANIM]**, **[READ]**

[ANIM], **[READ]**. If you are going to specify that if the only way a student can get an answer right is by exactly matching an expected answer, then of course you are only going to see answers that exactly match it.

Likewise, there are also some questions that are just not suited for this type of analysis, questions which are just too specific for students to give anything but the same answer (for it to be marked correct).

Future Studies

- 'Phrases' instead of single words.
- Extended responses from interviews.

So, where to for future research?

[ANIM] One direction could be to look at multiple words used together in sequence. Certain phrases may yield the student's thought processes behind their answers.

[ANIM] Another avenue would be to look at interviewing students, getting the student to explain their thought process when answering questions. Perhaps similar word usage will also be apparent in verbal communication as it was in the written when looking at lower and higher performing students (or students at different stages in their course).

[IF TIME]

Now, there is more to the second point than that what I have suggested. it has its basis in one of the papers we read which helped motivated this study.

Future Studies Motivation: Linguistic Enquiry and Word Count (LIWC)

- Korean students and experts were asked to recount their real life experiences when attending cultural events.
- Study showed that the words used in responses to questions could be accurately used to determine a students level of competence.
- An interview may be equivalent in length.

Kim, K., Bae, J., Nho, M.-W., & Lee, C. H. (2011). How do experts and novices differ? Relation versus attribute and thinking versus feeling in language use. *Psychology of Aesthetics, Creativity, and the Arts*, 5(4), 379-388.
<http://dx.doi.org/10.1037/a0024748>

LIWC, or Linguistic Enquiry and Word Count, is essentially a huge database of English words with tags to those words. For example, words such as 'happy', 'exuberant' and 'excited' might all have the tag 'positive'.

A Korean research team took the database and made their own version of it in Korean (called KLIWC). **[ANIM]** They then performed an experiment, they sent both students and experts to cultural events and then asked them to reflect on their experience at said event. It is important to note, that the experts and students they recruited were from different academic fields.

The result of their study was that **[ANIM]**, they identified that experts and novices differed in the frequency of certain words used. And that these patterns of words were domain independent.

Now, as you could imagine, there is quite a bit of difference between what they did and what we did. It is the point of a correct answer to an EiPE question to be short in length. Definitely much shorter when compared to the lengthy responses the Korean study received as reflections.

Although recounting your experience at a programming cultural event may provide a response of similar length, I doubt it will tell you much about how competent you are as

a programmer. **[ANIM]**, a transcribed interview, where a student reflects on their thought process when answering a given problem, could be of equivalent length to reflections of the KLIWC study. And patterns of words in that transcription may reveal more about the students level of development.

Questions?

Email us at:

Thomas Pelchen: Thomas.Pelchen@student.uts.edu.au

Raymond Lister: Raymond.Lister@uts.edu.au

So this is the end of the presentation. We go over all of the 12 questions in our paper, so I hope I have tempted you into looking those up.

Does anyone have any questions for me?

[If asked]

Average words: We ran ANOVA tests on the average length of each answer, apart from one question, the results were not significant.

Corrections: if needed, we ran a Fisher Exact test as a correction for 2x2 tables and a Fisher-Freeman-Halton exact test for 3x2 tables.