Some personal thoughts and comments on the New Zealand Scholarship Statistics Sample Examination and Schedule[1]

Mike Camden, Rachel Passmore, Ross Parsonage, Ruth Kaniuk, Dru Rose, Matt Regan

Scholarship statistics: opportunities for fresh directions (Mike Camden)

The new performance standard for scholarship statistics derives directly from the statistics strand of the curriculum. It does not derive from the achievement standards and it does not depend on the old scholarship Statistics and Modelling. Hence scholarship statistics is free to lead off into the future that the curriculum envisions. I'd like to see it take the directions noted below.

The scholarship examination needs to:

- Focus on statistical thinking, which is about variation in several forms. It can assess skills with these issues: What does the variation look like? Why is it like that? and maybe, What can we do about it? The Sample Examination takes this direction for the first issue.
- Cover all the Level 8 Learning Outcomes in the Statistics Strand. The Sample Examination does that well. It has time series, surveys, experiments, probability, inference with randomisation, with statistical literacy and the statistical enquiry cycle underlying most of it. The way it covers some of them still needs to be freed from the past.
- Use adult real contexts. The Sample Examination has much success here (as in the use of the Labour Force Survey), but there is room for further progress, particularly in probability.
- Show leadership in good statistical writing. The Sample Examination is often successful with respect to this issue, but some questions (Question 2 and Question 3) have obscure sections.
- Use actions that really do happen in statistical practice and avoid making up methods which statisticians would never use in practice.

We should take this rare opportunity, in the development of our new scholarship system, to move scholarship statistics in a direction which is consistent with 21st Century statistical practice.

In an attempt to illustrate what I mean with regard to the last bullet point above, I list and comment on the following 5 specific examples from the Sample Examination:

1. Fitting straight lines to smoothed series in Time Series (as in Question 1). We can now abandon this historical habit completely. We can provide a graph and the seasonal factors, ask for forecasts, and hope that students (in the absence of software) do a sensible pencil extrapolation of the smooth from the last few values. If we do give a regression line, our aim would be for students to say how irrelevant it is as a fit and as a part of forecasting.
2. Performing mental gymnastics with probability (Question 4). I ask whether anything like this happens in real statistical (probabilistic) practice, and whether it really is statistical thinking. Let's omit questions like this, and ask questions in real contexts, e.g. about relative risk. Or we could ask for a critique of how a probability model matches a real situation or real data.

[1] http://www.nzqa.govt.nz/qualifications-standards/awards/scholarship/scholarship-subjects/statistics/sample-resources/

3. Taking a sample from a sample (Question 2 – the Labour Force Survey Sample). In practice, a statistician might ignore the weights of the complex sample, treat it as a simple random sample, then do bootstrap samples of the whole thing. I don't think that a statistician would take a subsample as the question suggests. Statistical practice is quite complicated enough already!
4. The use of models in the bivariate situation (Question 3). It provides a relationship with subsets, non-linearities, and outliers; and it provides some models. That is a great start. However, the examination now needs to imitate statistical practice. The examination needs to say which records were used in each model, so that students can critique them and decide if they want to make use of any of them.
5. Using resampling methods and normal based distribution methods (Question 2). The arrival of resampling methods means that we can find a 21st Century balance between resampling and normal distribution based methods.

In these examples, the Scholarship Examination needs to provide the evidence that a student would get from the data with the help of software. The student can then assess this evidence and demonstrate statistical thinking. In some of these situations, a student could demonstrate fine statistical thinking skills, during the examination, with the provided graph and the use of a pencil.

Question 1 (Rachel Passmore)

The main focus of questions was on observation techniques – how much has a series gone up/down/stayed the same with some quantification.

I was disappointed that students were not asked to interpret **WHY** the patterns they observed might have occurred. Some would say this is not possible as different contexts will exclude certain students. However, we are talking about Scholarship students here so I don't think it is unreasonable to expect them to be able to make at least some inferences about employment and unemployment data.

Students are asked to make six distinct observations about the time series relating their comments to components of time series. I consequently found their choice of presenting seasonally adjusted data a little odd, as the seasonal component is often an important factor that warrants discussion. Only Figure 1 displays a series that is not seasonally adjusted but the scale makes seasonal variation almost impossible to detect.

Schedule – talks about 'significant December quarter seasonal effects' in Figure 1 - which there may be, but the scale of the graph does not really allow this statement to be made.

Time variable, question 1(b). Their time variable increment is 0.25, rather than the usual 1. Not sure why this additional level of complication was required.

Question (b) (ii) talks about the trend closest to the forecast period being different to the trend line. I would have expected some statement here as to whether this will result in over or under-prediction. It would also have been informative to have some indication of the level of accuracy of the forecasts, which with this traditional approach to forecast production is hard for students, but SO EASY with iNZight !

It is disappointing that we are still expecting students to extrapolate a linear trend when it is obvious to even the most novice of observers that this is not a sensible approach and is unlikely to lead to sensible predictions.

Question 2 (Michelle Dalrymple)

Overall, too much focus on normal-based methods for constructing CIs (use of the p +/- 1.96*sqrt(p(1-p)/n) formula) and not enough on assessing understanding of resampling methods.

The Working-age population count (result) of 3,448,000 (Figure 6) is not based on a sample of 29,456 people selected from the NZ working-age population. (See Paragraph 2, preamble.)

There are some numerical typos in the Question (Table 1) and in the schedule which can cause a bit of confusion.

Question 2a

This part seems to be straight from the old curriculum with the main part requiring the calculation of limits using Normal based theory. The only bit approaching scholarship standard is the determination of a sample size *n*.

The revised curriculum **S8-2** mentions "recognising the relevance of CLT", but the second tier information (as linked in the scholarship performance standard) states *"The approach to teaching CIs may be entirely based on simulation using technology"* and under What is new/changed *"Resampling and randomisation methods will be used to generate CIs and to assess the strength of evidence; this means that the central limit theorem is de-emphasised as a basis for CIs. Rather the focus is on the logic behind the inference".*

Does this Sample Scholarship Exam question signal that teaching a normality-based approach to constructing CIs is required despite the above statements?

Question 2a(i)

Calculating the limits for the CI would have been a standard external NCEA question under the old curriculum.

For finding *n*, students need to know to take the working-age population 3,448,000 as the true value. The wording in the preamble could be misleading and/or confusing. (See statement above)

Question 2a(ii)

This question is not clear (until you look at the schedule) as to whether it's looking for some interpretation of the CI limits found in part (a) (expected response: *With 95% confidence, the true value is estimated to be somewhere between x and y*) or whether it's an explanation of the 95% confidence level (expected response: *This process captures the unknown value 95% of the time* –words/ideas to that effect).

In either case it can be pretty much a learned response which can be easily regurgitated without any demonstration of understanding.

Evidence 2a(ii):

If 'the interval' referred to in the first bullet is (0.929, 0.936) then the response given in the first bullet is just as unacceptable as the not acceptable response given in Note 1. (Making the statement about the probability of a variable (the interval) and not about the constant value (the parameter value) is not the issue here.)

Question 2b

Table 1 shows the cumulative percentage distribution of the *resample percentage.* Use of the term 'percentiles' may be more familiar to the students.

Question 2b(i)

What's the point of the required linear interpolation?

The schedule doesn't actually answer the question, but just gave a CI for the percentage in the labour force (rather than number).

Question 2b(ii)

Initial confusion as to whether the previous survey was the previous quarter survey used in part (i).

Evidence 2b(ii):

I wouldn't have realised without looking at the schedule that such an in-depth discussion of sampling error was an expected response. If it was expected, then perhaps a discussion of the sampling error associated with *the difference* between the two survey percentages would have been appropriate at this level.

It appears that the expected response should be centred on the idea that point estimates don't take sampling variability into account and hence we should not use the direction of the estimate of the difference to make a claim about the direction of the true (population/underlying) difference. This idea is a critical one when considering the difference between two population parameters but is hardly one that should be the essence of a question at scholarship level. Taking account of sampling idea is a big idea which is now pushed in earlier years but it will take a year or two before those students reach this level and so maybe assessment of this idea is ok at this level for now.

Re: Note 1 If the previous estimate had been, say 94.00%, then should the following argument be accepted? Since the *previous* estimate is in the *new* CI (93.085%, 95.205%) then the two CIs for the two survey percentages will very much overlap (more than one half of one interval overlaps the other) and hence there is no evidence that the new true population percentage is greater than the previous true population percentage. Would we expect students at this level to know that overlapping 95% CI for the individual population percentages doesn't necessarily mean that the difference between the two survey percentages is nonsignificant (at the 5% level)?

Question 3 (Ross Parsonage)

Comments on how the question reflects the spirit of AS 3.9

This question is very similar to the original question in the 2011 paper and has therefore missed an opportunity to reflect important differences between the old Achievement Standard 3.5 and the new Achievement Standard 3.9.

The most important weakness in this question is that the question is not constructed in a way that poses an appropriate relationship question (see Explanatory Note 3, first bullet point).

A second weakness is that the question does not provide a multivariate dataset (see Explanatory Note 3, first bullet point).

A third weakness is that the amount of contextual information is limited.

I appreciate that some elements of the new AS 3.9, which is internally assessed, can’t be incorporated in an externally assessed question but I wish the question had provided a multivariate dataset (of 3 or 4 variables) and that more contextual information had been provided. The elements of the statistical enquiry cycle that are difficult to assess externally are (from Explanatory Note 3):

- Posing an appropriate relationship question
- Selecting and using appropriate displays
- Finding an appropriate model

Candidates should not be expected to find an appropriate model but they should be required to discuss the appropriateness of given models (as is done in this question).

Scholarship candidates should be investigating bivariate measurement data, with justification and with statistical insight. Two groups are evident in the scatter plot, along with an outlier. The candidates cannot use contextual reasons for explaining the presence of the two groups or for the presence of the outlier. If values of a third or fourth variable had been provided then candidates could have used these variables to provide contextual justifications and giving opportunities to integrate statistical and contextual knowledge.

I feel that this question, as written, does not provide many opportunities for candidates to show statistical insight.

Seven comments about the question itself

I don’t like (b), especially the introduction and part (i).

Comment 1

Is the description of E too vague?

Is it the time the person has been employed as a teller at that bank?

Is it the time the person has been employed as a teller at any bank?

Comment 2 (Part (b) introduction)

The regression equations are expressed in terms of y and x and would be better expressed using S and E.

Comment 3 (Part (b) introduction)

Which points have been used to produce the three linear models? It is not clear and, in my opinion, it is not obvious. There are more than three possible options for fitting models. I can find five without even considering whether (19, 26.4) is in the "upper" group or the "lower" group.

1. Use all 36 points
2. Use 35 points (without the outlier (15, 5.1))
3. Use the upper group of 10 points
4. Use the lower group of 26 points including the outlier
5. Use the lower group of 25 points without the outlier (15, 5.1)

I think it is unfair to expect candidates to "guess" which points have been used. Even if an extra variable (or two) had been provided to help justify the different groups it does not help the candidates know the points used to produce each regression line.

In fact:

- The regression equation y = 0.411x + 15.1 is obtained using all 36 points
- The regression equation y = 0.255x + 23.6 is obtained using the 10 points in the upper group.
- The regression equation y = 0.646x + 9.84 is obtained using the 26 points in the lower group (with the outlier included)

Comment 4 (Part (b) (i))

This is related to comment 3 above but includes an extra concern.

How can candidates justify their selection if they are not certain of the points used to produce the fitted line?

The extra concern is that in AS 3.9 students are expected to use visual inspection. When candidates are justifying their selection of the line I would expect them to comment on the appropriateness of their chosen line. How can they do this without the lines being drawn on the scatter plot? I would not expect the candidates to have to plot the line(s).

Comment 5 (Amount of scaffolding)

Is there too much scaffolding? See (b) (iii) and the bullet points in (c).

Comment 6 (Part (b) (iii))

In (b) (iii) is "validity" the best word?

I think validity means "Does it measure what it claims to measure?" In bivariate analysis this is about how well the assumptions of the analysis are met. This content is beyond secondary school level. I think precision is a better word because it can be related to the degree of scatter about the fitted model.

Comment 7 (Part (c))

As part of showing statistical insight candidates are encouraged to consider other relevant variables (see Explanatory Note 2). Part (c) suggests three variables. Candidates could be asked to identify some more variables that would help answer a posed relationship question.

Comments on Assessment Schedule

Part (a)

The answers to (a) do not reflect the changes between the old AS 3.5 and the new AS 3.9. Comments on data features should refer to the graph and be contextual. Use of symbols for variables is not sufficiently contextual and terms such as "positive" and "correlation" are similarly not referring specifically to visual aspects and are not sufficiently contextual.

Consideration for O and S should reflect the meaning of "justification" and "insight" as described in AS 3.9. The current assessment schedule tends to count the number of observations made and does not take account of how well the candidate uses justification and statistical insight.

The answer to (a) mixes comments on data features with comments on selecting appropriate models. The way the question is structured I would not expect students to comment on choosing models in (a).

Here is an answer to (a) that is more consistent with the NZQA exemplars for AS 3.9

One teller, the one with 15 months experience who serves 5.1 customers per hour, on average, can be regarded as an outlier.

There appear to be two subgroups; an "upper" group of 10 tellers [(1, 23.8), (4, 23.0), (3, 25.6), (13, 28.3), (6, 24.6), (8, 25.1), (10, 26.2), (14, 28.3), (16, 28.5), (19, 26.4)} and a "lower" group of the other 26 tellers. The teller with 19 months experience and a service rate of 26.4 customers per hour, on average, could be in either group.

Using the whole group of 36 tellers, tellers with less experience tend to serve fewer customers per hour, on average, and tellers with more experience tend to serve more customers per hour, on average. The relationship between experience and service rate appears to be non-linear with the increase in service rate being reasonably constant from one month of experience to about 12 months of experience but from 12 months to 20 months of experience the rate of increase in service rate gets lower. There is quite a lot of scatter about the trend so the relationship can be described as moderate.

Considering the "upper" group, tellers with less experience tend to serve fewer customers per hour, on average, and tellers with more experience tend to serve more customers per hour, on average. There appears to be a linear relationship between experience and service rate for this group with little scatter about a linear trend so the relationship is strong.

Considering the "lower" group, tellers with less experience tend to serve fewer customers per hour, on average, and tellers with more experience tend to serve more customers per hour, on average. The relationship between experience and service rate appears to be non-linear with the increase in service rate being reasonably constant from one month of experience to about 12 months of experience but from 12 months to 20 months of experience the rate of increase in service rate levels off. There is not a lot of scatter about the trend, but slightly more than in the "upper" group, so the relationship can be described as quite strong.

Part (b)

I'm not commenting on (b) (i) and (b) (ii) because I think the question is too flawed.

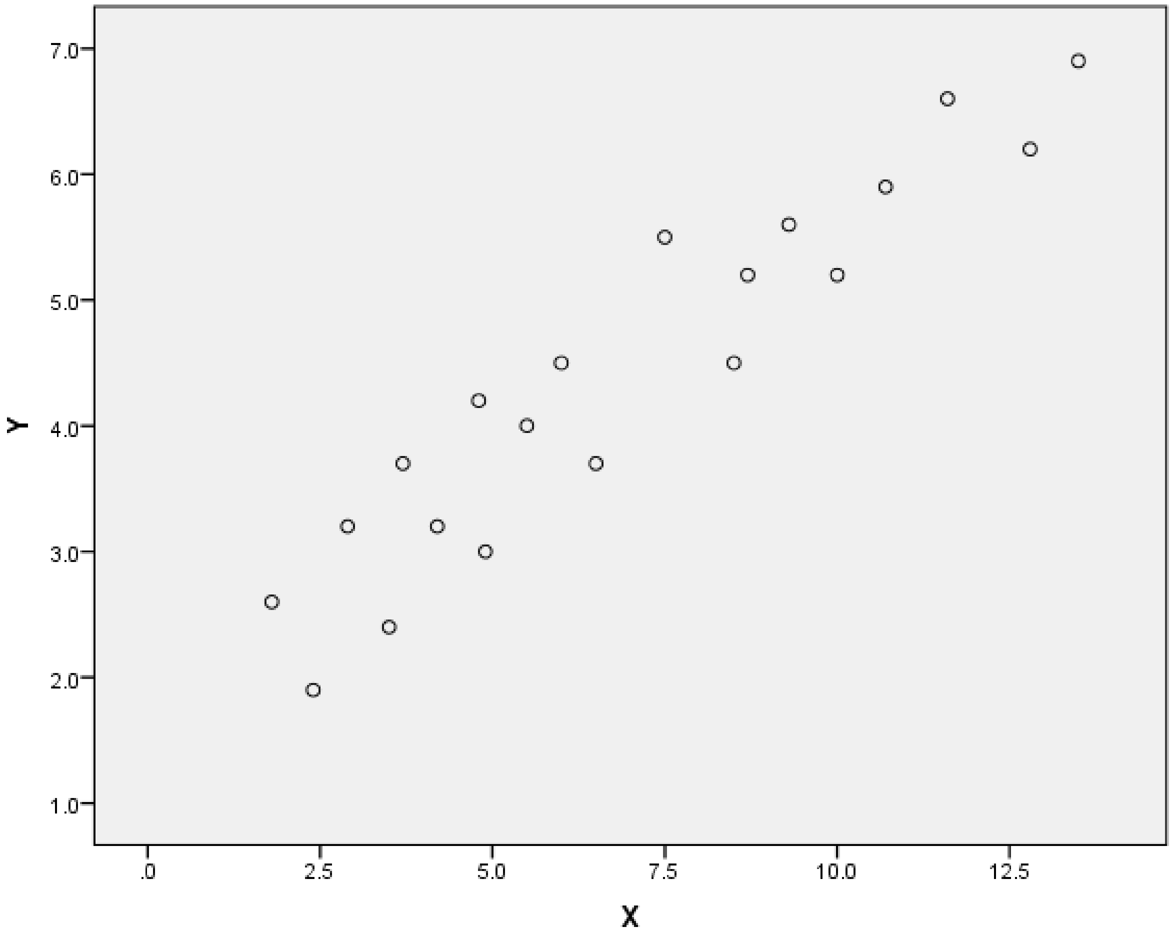
In (b) (iii) I strongly disagree with the justification for the predicted value of S for E = 9. Comparing the predicted value of S with observed values for S (for the same or similar values of E) does not justify the validity of a prediction. It also raises (again) the issue of what does validity of a prediction really mean.

Here is an example to illustrate the flaw in the argument provided in the answer.

The equation of the regression line fitted to the scatter plot below is y = 1.722 + 0.386x

For x = 7.5, the predicted value of y is 4.62

The only observation at this part of the x-axis is (7.5, 5.5). On the scale of these values the observed value of y is well above the predicted value. Does this make the prediction invalid? I don't think so.



Part (c)

I agree with Ruth's comments here (as I do in most places). I would not expect the candidates to refer to a correlation coefficient. A description of the strength and direction was asked for and that is what candidates are likely to provide.

Question 3 (Ruth Kaniuk)

The question seems to be little changed from the original, but then the bivariate topic possibly has less change than many others.

I think that the question is appropriate for scholarship level. The question is largely OK for the new standard. The word 'validity' in b(iii) might be better replaced with 'precision'? b(i) could include the direction to explain what each of the given lines model. In (c) the suggestion that 'you may find it helpful to use diagrams' is unnecessary scaffolding. 'use appropriate statistical terminology' seems to have been included to create a dubious distinction between O/S.

The published answers are such that I doubt that candidates would achieve a 'schol' and markers would be doing a lot of remarking.

Also, why does each <u>part</u> of a question have an O/S/P- that would be a huge jump in standard expected. To date there has been one O per full question. Some of the sample answers appear to be in 'adult speak' and seem to be included to create an 'artificial' distinction between O/S

Question 3(a)

Bullet point 1 – really has two distinct ideas in it: group and outlier- why lump them together?

Bullet point 5 – quite pedantic about regression as opposed to association/ relationship. The question talked of relationship. Experience tells me that students – even our better ones – would not give an answer remotely like the one supplied.

I would think that students should be expected to:

- identify the outlier and give values of the point
- identify the groups
- suggest that the relationship for the upper group is that tellers with more experience tend to serve more customers per hour on average, (positive)
- that there is a gradual, reasonably constant, increase in S, so the relationship could be modelled with a line
- the relationship is quite strong because of little scatter
- suggest that the relationship for the lower group is that tellers with more experience also tend to serve more customers per hour on average (positive)
- the increase in S is quite rapid for values of E between 2 and 10 and then flattens off so the relationship could be modelled by a curve
- there appears to be slightly more scatter for these points than the ones in the upper group

Quesiton 3(b)

Given solution seems reasonable, but I would hope that a prediction interval might be given. Vague and possibly indistinguishable difference for O/S

Question 3(c)

I would not expect students to estimate/comment on correlation coefficients. They should comment on strength and direction.

Question 4 (See Mike Camden's comment above.)

Question 5(a) (Dru Rose)

I think the question set was a fair question with plenty of scope for students to discuss a variety of aspects: with the ability to comment on positive as well as negative aspects of the survey methodology.

The provided report is about the right length and used an accessible context.

I think the first bullet point in the question (Evaluate the statistical evidence and processes) is a vague instruction for students and better guidance on what exactly that means is needed, e. g. 'Evaluate the validity of the survey method used' or even more generally, 'Discuss the reported findings in relation to the intended purpose of the survey.'

However, the students used as guinea pigs seemed to have understood what was required and the remaining bullet points gave clear pointers on what to include.

The students' answers raised valid points missing from the assessment schedule:
People using internet banking have ready access to their account balances and so would probably have accurately reported the balance in their savings accounts (disputed in the assessment schedule) but may well have not known exactly how much was invested in high risk shares -especially if they were in a managed fund and not actively buying and selling shares themselves.

One student estimated the margin of error for a difference in proportions of men and women investing in high risk shares (1.5 x $1/\sqrt{600}$) and commented that the %difference in the survey is larger than this, lending some weight to the validity of the claim. However the findings should not have been transferred to all NZ from a sample of only XYZ bank customers.
This student also picked up that age, education and income level are likely to affect the amount a person saves and type of savings and these were asked in the survey but not reported on.

Random selection of the participants should have ensured a reasonable cross-section of people across age-groups, income brackets etc. (not mentioned in the schedule).

Question 5(b) (Matt Regan)

Overall a reasonable question for scholarship mainly because the two outputs give inconsistent results. The question provides the opportunity to differentiate between Scholarship Performance and Outstanding Performance as defined in the Statistics Scholarship Performance Standard.

A major concern that I have with this question is in the schedule particularly for part b(i). Some good teaching and learning issues are provided by the given schedule including a common misconception that we should try to avoid passing on to our students.

I appreciate the schedule appears in note form and may not necessarily contain a complete list of acceptable evidence or all the issues that may arise in student responses.

Question 5(b)

I would not have described the re-randomisation distribution as a bar graph nor a histogram but *a dot plot* of the differences between the re-randomised medians or means.

Evidence 5b(i):

Real care needs to be taken with the wording in statements similar to that of the bracketed statement in the first bullet – a difference of 2.2 or more is likely to occur *by* chance and *simply to be due to random allocation to the two groups*. It's the word 'by' and the second half (italicised) which cause the problem. (A similar statement is made in the second bullet when dealing with the difference between the means but that appears to be a cut and paste typo.) This statement says that the source of this observed difference is likely to be the random allocation (or chance) which is a very common but totally incorrect inference to make from this output. We need to be aware of this common misconception and almost explicitly draw it to our students' attention – a large tail proportion does NOT allow us to infer that the observed difference is likely to be due to the random allocation (chance). With a large tail proportion the jury remains out – the observed difference could be either due to chance alone or it could be due to the treatment together with an element of chance, we are not able to make a claim as to which.

To me, an unacceptable omission in the evidence list is the overall answer to the question 'Is the treatment (the training course) effective?'. The schedule simply interprets the outputs in Figures 8 and 9 separately, and thereby avoids having to deal with the inconsistency in their outputs.

There should be a clear statement in the conclusion as to whom the findings of the experiment apply. Since we are not told how the 40 bank tellers in this experiment were selected then strictly speaking our findings about the effectiveness of the training course only applies to these 40 tellers and we should be very careful when making statements which generalise the findings to any wider group of tellers. (This is mentioned in the schedule under *Evidence (ii)* but it should really be part of the conclusion statement.)

Another teaching /language issue is raised in the schedule by the use of the word 'real' in the phrase: "a real difference between the median time taken to complete a particular transaction by group A and median time taken to complete a transaction by group B." What is meant by the use of 'real' in this context? Should it's use be postponed until students have had some exposure to formal significance testing and its associated language and terminology.

It could also be appropriate to briefly mention in the conclusion the estimated size of the training course effect – is it worth the cost of a training course? That is, even if, on average, the training programme makes a difference, is the difference big enough to warrant the training course? (The practical significance versus statistical significance issue.)

Evidence 5b(ii):

I don't think that the factors listed in Bullet 2 (group size) and Bullet 3 (selection of study participants) should gain any credit in the answer to this question.

Bullet 2: Group size is a factor to consider in the design of an experiment but I don't agree that it is a factor which needs to be considered in order to make a claim about a causal relationship. We can have well designed experiments with much smaller group sizes than in this study in which we could validly make a causal relationship claim. It is true that with random allocation with smaller group sizes there is the potential for greater non-comparability between the groups but this risk of greater imbalance is not grounds to question the validity of any resulting causal claim. It's the chance mechanism (random allocation) within the study design (the basis of the randomisation test) which allows the test procedure to handle any imbalance between the groups.

Bullet 3: It is possible to make a claim of a causal relationship between treatment and time taken to make a transaction without considering the selection process of the 40 bank tellers in this experiment but, in doing so, that claim would only apply to the 40 tellers in the study. If we then want to make a sample-to-population inference (i.e., generalise the causal relationship to the population of all tellers at XYZ bank) we would need to have ensured that the 40 participants in the study were a random sample from this wider population.

The key factor in making a causal relationship claim is that the study is a well designed and carefully executed experiment – well designed means that efforts are made to make the comparison between the two treatment groups fair. 'Fair' in a comparable sense, that is, the groups are as similar as possible in all respects except that each group receives a different 'treatment'. *Playing fair* is the overriding principle that we need to consider in designing a study like this one.

In this study random assignment of the 40 bank tellers to one of the two treatment groups is one mechanism that is used to attempt to make the groups balanced with respect to other factors which may affect the time taken to complete a transaction (e.g., age, experience etc). We can't guarantee that the random assignment will exactly balance the groups but the random assignment does allow us to classify all explanations for the observed difference, other than the difference-in-treatment explanation, as *chance explanations*.

Another factor which may affect the comparability of the two groups is that one group of tellers receives some special attention (a training programme) and the other group doesn't receive special attention and perhaps it's this special attention, rather than the programme itself, which is the cause of this observed difference in transaction times between the two groups. One way of nullifying this difference would be to try give the control group some special attention programme as well but nothing to do with the transaction times, e.g., maybe a programme on personal appearance training. As ethically as possible, keep the study participants in the dark as much as possible! Perhaps, don't let the 40 tellers know that they are part of an experiment; try not to tell the tellers in the treatment group that the goal of the training is to speed up the time to complete a transaction; etc. (I am trying to think about factors which can produce Hawthorne and placebo-type effects here and that it would be reasonable for students to mention without knowing the technical terminology/names.)

Consideration needs to be given to when/how the transaction times are taken. Again there is a need to try to play fair, e.g., make sure that the time of day that these transaction completion times are taken is comparable for both groups.