No part of the candidate evidence in this exemplar material may be presented in an external assessment for the New Zealand Scholarship award.

93201A

932011

S

SUPERVISOR'S USE ONLY

TOP SCHOLAR

NZQA
NEW ZEALAND QUALIFICATIONS AUTHORITY
MANA TOHU MĀTAURANGA O AOTEAROA

QUALIFY FOR THE FUTURE WORLD
KIA NOHO TAKATŪ KI TŌ ĀMUA AO!

Tick this box if there is no writing in this booklet

Scholarship 2020 Statistics

2.00 p.m. Friday 20 November 2020
Time allowed: Three hours
Total score: 40

ANSWER BOOKLET

Check that the National Student Number (NSN) on your admission slip is the same as the number at the top of this page.

Write all your answers in this booklet.

Show ALL working. Start your answer to each question on a new page. Clearly number each question.

Check that this booklet has pages 2–24 in the correct order and that none of these pages is blank.

YOU MUST HAND THIS BOOKLET TO THE SUPERVISOR AT THE END OF THE EXAMINATION.

Question	Score
ONE	
TWO	
THREE	
FOUR	
FIVE	
TOTAL	/40

ASSESSOR'S USE ONLY

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

1a(i) The relationship between the numbers of words and number of pages for 65 of the most popular fiction books written in English is a positive, linear association (trend). The positive association is relatively strong with fairly constant scatter. ~~either side of the trend~~. There are no obvious groupings in the data however there are 2 outliers – one at approximately (2000, 525) and one at (1900, 150).

1a(ii) The number of pages of a book might not precisely predict the number of words in the book since different books ~~uses~~ use different font sizes. If a book uses a large font size, it may contain less words than a book with the same number of pages but a smaller font size, due to large fonts occupying more space/pages.
Furthermore, a book may contain images which take up space and pages. If a book has images, it may contain less words than a book without images but the same number of pages.
Hence, without considering these factors, ~~the~~ it is hard to precisely predict word count from pages.

1a(iii) Since the number of words ~~in the~~ is measured in thousands while number of pages is not, the gradient of the model of 0.2202 indicates that on average, the number of words ~~in the~~ per page is 220 (0.2202 × 1000). Hence, according to this model, the number of words in a book increases by 220 words per page. The model has a y-intercept of 57.84 ~~that probably accounts for~~ when a y-intercept of 0 would be expected (since books with no pages should have no words). However, relative to the total number of words in a book, 57.84 is very small so can be ignored. ~~Though~~ 220 words and 250 are different by 30 words which is a 12% difference. This could be considered significant enough to say that the statement that 'for fiction books there are, on average, 250 words per page' is unsuitable. ~~Instead, it should say that the~~ * see page 17 → seen

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

1 ~~An average 220 words per page (for fiction books).~~

1b(i) A claim made in the report is that the difference in percentage of NZ adults who had read or started to read at least 1 book in the past year is not statistically significant. This claim is based on the percentages of 86% of NZ adults reading in March 2018, compared to the slightly higher percentage of 88% in March 2017.
The point difference of these percentages is 86 − 88 = −2% (2018 lower than 2017).
The MOE for this difference:

$2018\ MOE = \frac{1}{\sqrt{2261}} \times 100 = 2.10\%$

$2017\ MOE = \frac{1}{\sqrt{2082}} \times 100 = 2.19\%$

$Avg = \frac{2.10 + 2.19}{2} = 2.15\%$

$1.5 \times Avg = 1.5 \times 2.15 = 3.22\%$

$CI = [-5.22\%, 1.22\%]$

Therefore, with at least 95% confidence, it can be inferred that the percentage of NZ adults who had read or started to read at least one book in the past year as at March 2018 is somewhere between 5.22% less and 1.22% more than the percentage of NZ adults as at March 2017. As this confidence interval ~~is~~ contains 0 (is not entirely negative), a conclusive claim cannot be made that a lower percentage of NZ adults read in the past year ~~not~~ as at March 2018 than March 2017. Hence, the claim that the difference is not statistically significant is valid. The survey percentages of 86% and 88% are outside the 30-70% range which the 'rule of thumb' applies to meaning that the confidence interval calculated is overestimated ~~as~~ (i.e. greater than 95% confidence). However, even ~~the~~ the true 95% confidence interval would likely still include 0 so the claim stands.

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

11b(ii) A non-sampling error associated with being asked how many books someone read this year is that people may feel ashamed if they hadn't read any books since it could make them feel like they were lazy if they ~~did~~ didn't. ★ This could cause people ~~to~~ who didn't read, to lie about having read a book to appear more diligent. This could result in the 86% ~~&~~ survey percentage being an overestimate of the percentage of NZ adults who had read/started to read at least one book in the past year. ie. the actual percentage could be lower. Another non-sampling error is that for people who read a lot of books, they may not have kept track of exactly how many books they read in the past year. Therefore, when asked how many books they read in the past year, they may just ~~guessed~~ provide a guess of how many. Therefore, some responses may be an inaccurate reflection of how many books someone actually read as it is hard to remember exactly how many. Therefore, the estimate that on average ~~#~~ 35 books per reader were read could be either over or underestimated as people could overstate the number of books, or understate if they forgot how many. Unless someone writes down how many books they read, the ~~most~~ chances of someone accurately answering the question either by accident or because they don't know, is high.

★ This is because reading is generally seen as something everyone should do to improve their breadth of knowledge & vocabulary.

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

2a(i) ~~##~~ Using the bootstrap confidence interval of (108, 143), it can be inferred with 95% confidence that ~~books on Project Gutenberg were published somewhere between 108 and 143 years prior to~~ the median age of all books on Project Gutenberg is somewhere between 108 and 143 years old ie. books were published between 108 and 143 years prior to 2020. This means that with 95% confidence, it can be said that the median publishing date of books on Project Gutenberg is somewhere between 1877 - 1912 (2020-143, 2020-108). As this entire range/interval is before 1924, Project Gutenberg's claim that most of their eBooks were published before 1924 is valid/supported. 95% of their books were published between 1877 - 1912.

2a(ii) From the sample data distributions it can be seen that there is slightly greater variation in the mean age of books on LibriVox than Project Gutenberg. This is because the mean age of books in the LibriVox sample ranges from 25 - 420 years while the mean age of books in the Project Gutenberg sample ranges from 5 - 240 years. However, this difference in variation may be due to the small sample size of each (80 and 200).

It can also be concluded that the mean age of books in the LibriVox sample is ~~216~~ 458 years older than the mean age of books in the Project Gutenberg sample.

From the bootstrap confidence interval, with 95% confidence it can be inferred that back in the population of all books of each library, the mean age of LibriVox books is somewhere between 14.10 years younger and 22.76 years older than the mean age of Project Gutenberg books. Since this confidence interval includes ~~&~~ zero (not entirely positive), it cannot be concluded that ~~in~~ LibriVox books are older than Project Gutenberg books. Since zero lies close to the middle of the confidence interval, it is likely that another bootstrap confidence interval using different samples would also contain zero. Hence, no significant difference in the age of books on each library can be concluded.

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

2a(iii) Binomial distribution:

This distribution can be used since there are a fixed number of trials - 24 books in the sample. There are 2 possible outcomes - written in English (written in English) or not written in English. ~~Thes~~ A constant probability of success is assumed: $P(\text{written in English}) = \frac{19}{24} = 0.7917$. The event of one book being written in English is independent of whether another is written in English as books are published/~~wrote~~ written independently of each other (excluding sequels which we assume are ignored).

$p = 0.7917$ ~~n=24~~ $n = 24$

$P(X > 12) = 1 - P(X \leq 12) = 1 - 0.001428 = 0.9986$

~~99.9% of books available from Project Gutenberg~~

∴ There is a 99.9% chance that more than half of the books available from Project Gutenberg are written in English. Since this percentage is very large (almost 100) ~~the~~ there is sufficient evidence to conclude that more than half of the books ~~are~~ on Project Gutenberg are written in English.

Also, using $n = 60000$ ~~and inverse binomial distribution:~~

~~$P(X \geq 30000) = 1 - P(X < 30000) = 1 -$~~

$X = 47500$ i.e. $\frac{47500}{60000}$ books are expected to be written in English

~~$\text{range} = \frac{1}{\sqrt{n}} = \frac{1}{\sqrt{24}}$~~ ∴ more than half

2b The problem being investigated is ~~about~~ how much longer a person who arrives at the library in the morning stays, than a person who arrives in the afternoon.

The planning step would involve planning the days which people's (arrivals/staying time) will be observed at the local library. Each day should be observed so that any differences in staying time between different days of the week can be observed. Also, what classifies as morning and afternoon must be established eg. before 12pm (for morning).

The next step is data collection. This could ~~either~~ be carried out by manual

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

observation but a more effective method may be getting people to sign in when they arrive & sign out when they leave. This saves the need to watch people. Once data has been collected over several days and weeks to give a representative sample, the data must be analysed. ~~A paint stuff the same~~ A bootstrap analysis should be carried out to determine the ~~significance~~ size of the difference in staying time between morning & afternoon arrivals. The samples of morning and afternoon arrivals should each be sampled with replacement to produce ~~so~~ bootstrap samples of the same size of the original samples. Then, the difference in mean staying time should be calculated between the two resamples, and plotted into a bootstrap distribution. This process should be carried out many times to produce a 95% bootstrap confidence interval, between which the difference in staying time between morning & afternoon arrivers is likely to lie. Firstly, if the entire confidence interval is positive, it can be concluded that people who arrive in the morning stay for longer than those who arrive in the afternoon. Then, the confidence interval values tell us ~~what the~~ the range of how much longer they stay, with 95% confidence.

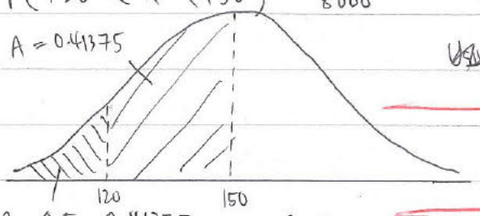
QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

3a) For both fiction and non-fiction books, reading activity peaks around 10 pm to 11 pm each day, with the exception of non-fiction book reading activity also peaking at 11 am. For both genres, reading activity is also a minimum at 6-7 am in a day.

At all times of day, the reading activity for both genres is relatively similar except at 11am to 12pm when the reading activity of non-fiction books is much greater than fiction books. At 11 am, 5.00% of reading activity is for fiction books while 6.00% is non-fiction. This 1% difference is significant compared to other times of day where the difference is much smaller eg. 0.1% difference at 8pm. Therefore, at 11 am, if someone is reading they are 1.2 times as likely to be reading a non-fiction book than a fiction book. Although the other differences at other times are smaller, at 3pm onwards until midnight, reading activity for fiction books is higher than for non-fiction. This is likely because people like to read fiction for leisure after school/work while non-fiction tends to be read during the day for study.

3b) $\mu = 150$ normal distribution due to bell-shape.

$P(120 < X < 150) = \frac{3310}{8000} = 0.41375$



$A = 0.5 - 0.41375 = 0.08625$

Using standard normal distribution $\sigma = 1$ $\mu = 0$

$z = -1.364$

$-1.364 = \frac{120 - 150}{\sigma}$ $\sigma = \frac{-30}{-1.364} = 21.99 = 22$

∴ normal dist. parameters: $\mu = 150$ $\sigma = 22$

middle 95% of reading speeds is within 2 s.d. either side of mean

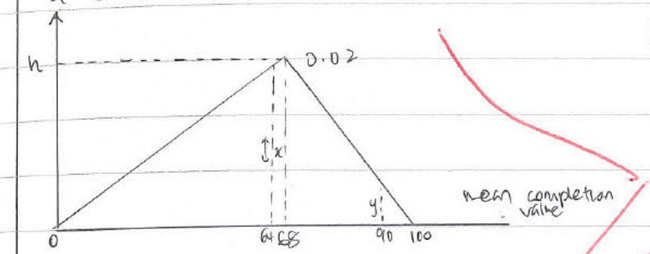
$150 - 22 \times 2 = 106$ $150 + 22 \times 2 = 194$

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

The lower and upper limits for the middle 95% of reading speeds for people in the study is 106 wpm and 194 wpm.

3c(i) Triangular distribution parameters:

$a = 0$ $b = 100$ $c = 68$



$h = \frac{2}{100} = 0.02$

$\frac{64}{68} = \frac{x}{0.02}$ $x = 0.01882$

$P(X < 64) = \frac{1}{2} \times 64 \times 0.01882 = 0.6024 = 60\%$

$\frac{10}{32} = \frac{y}{0.02}$ $y = 0.00625$

$P(X > 90) = \frac{1}{2} \times 10 \times 0.00625 = 0.03125 = 3.1\%$

A triangular distribution with the parameters min = 0, max = 100, mode = 68 would not be a good model for the mean percentage completion values for books available from the company. This is because in the study, about half (50%) of the books had a mean completion value of 64% or less. However, using a triangular distribution, 60% of books are predicted to have a completion value ≤64%. This is a 20% difference (10 percentage point difference) which is significant ie. the proportions are not similar. In the study, 5% of books had a mean completion above 90% while in the triangular distribution, 3.1% have a mean completion above 90%. This is a 1.9 percentage point difference which is not too large. However, due to the large difference in predicted proportion of books with a mean completion of 64% or less, a triangular distribution is not a good fit for the data.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

3b(ii). One factor that might explain the variation in mean percentage completion values for the books is the genre and type of book. Non-fiction ~~books are often~~ books such as encyclopaedias are often used for reference/research rather than normal reading purposes. Compared to fiction books which are more commonly read for leisure and thus read to completion, non-fiction books are more likely to only be partially read. Therefore, the books with relatively low mean completion values are likely to be non-fiction/research books. Meanwhile, ~~more~~ it is likely that a book read to completion (or majority read) is a fiction book. Hence, the purpose of that certain books are read for could explain the variation.

Another factor that might explain the variation is that different people have different levels of interest in reading and different amounts of time to read. Someone who is very interested in reading is probably more likely to finish or read more of a book they are reading than someone who doesn't like reading or doesn't have much time for it due to work or school. Therefore, depending on the kind of people who read certain books, the mean completion values for these books will be affected. Books read by a high proportion of people who are busy or don't like reading will tend to have lower mean completion values than books read by a high proportion of people interested in reading.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

4a) This experiment involves two independent groups being compared. One group is given text in serif font, and the other group is given text in sans-serif font to read, and their reading speed in number of words per minute is observed. Therefore, the explanatory variable is font (serif or sans serif) while the response variable is reading speed. The sample size is 238 students so each group has 119 students. The treatment was the 'font'. Students were randomly allocated into their respective groups to reduce bias ~~due~~ due to factors such as innate reading ability ie. people's abilities are evenly distributed between the 2 groups. The controlled variables were the text given (258 words) and the amount of time observed (one minute). The ~~sample~~ sample was taken from a population of university medical students who self-selected to participate.

4b) From the sample data, it can be seen that the sample of people who read sans serif text ~~read~~ had a mean reading speed of 1.75 words per minute faster than the sample who read serif font. The randomisation test output thus shows that a difference in reading speed of 1.75 words per minute ~~could occur due to~~ has a 40% chance of occurring due to chance alone. 40% is above the threshold of 10% meaning that the difference in mean reading speed is not statistically significant. It is unlikely that ~~any other~~ a factor other than chance (ie. font) was the cause of the observed difference in reading speed. The difference could have just been due to sampling variation. Therefore, it cannot be concluded that SansSerif causes ~~it faster for~~ people to read faster (by 1.75 wpm) than if they read serif font.

This result (that the difference in reading speed was likely not due to font) could have been expected since the study was carried out on university medical students. Medical students are used to reading a wide variety of text so are unlikely to be affected by a small change in font. They are probably well adapted

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

to reading a lot of different fonts and words due to their studies. Hence, comparing the reading speeds of medical students when given different fonts is unlikely to show any significant difference. If the study was carried out on people who aren't expected to read so much ~~conte~~ text usually, a more significant difference may have been observed.

4c(i) A much larges proportion of students in the study were female than male, with 80% being female and 20% male. For both specialisations, a higher proportion of students undertaking that specialisation are female with about 80% of both paediatric / therapy students being female and about 20% being male. However, a higher proportion of all students specialise in therapy than paediatrics, with almost 60% in ~~the~~ therapy and just above 40% in paediatrics. Students specialising in therapy tend to be younger than students specialising in paediatrics with a median age of 21 years old for ~~the~~ therapy and 22 for paediatrics.

(ii) The design of the experiment could be modified by recruiting more males to participate. This is ~~the~~ because there is a significant gender imbalance in the current study with ~~300% more~~ 4 times as many females than males. This may have caused female responses to font to be overrepresented, and male responses to be underrepresented. Males may show a different response when given different fonts to read than females however this difference would not be obvious due to the small proportion of males studied. The study also only covers ~~2~~ students in 2 degree specialisations - paediatrics & therapy. There are many more specialties than this so this sample in the study is unlikely to be representative of all medical students. Hence, students from other specialties such as neurology or surgery should be recruited. People specialising in different fields may also show different responses to fonts than those initially studied - so they should be included.

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

4d) prior to knowing this information, the results of the study were already not very generalisable due to being conducted on medical students who probably have better and faster reading abilities than the average person. However, the fact that the study was carried out using Cyrillic alphabet on Russian students who were fluent speakers of the language and had normal vision, further reduces the generalisability. Furthermore, the text used was on the history of medicine in Russia. The medical students will then likely be familiar with the content of the text so will be more likely to read it faster.

Serif / sans serif font may firstly affect the legibility of ~~font~~ Russian alphabet than English alphabet, having little effect on the ease with which it is read. Secondly, the fact that medical students would have been familiar with the content of the text means they probably would read it fast regardless of font. This prevents the ability to conclude the effects of the font. The students also all had normal or corrected vision which ~~most~~ some people in their day to day life are unlikely to have, since not everyone wears their corrective lenses at all times. Therefore, the very specific ~~example~~ demographic of the sample used in this experiment ~~let~~ significantly limits the ability to extend the results to anyone who does not fit the criteria of the participants, which is most normal people in the rest of Russia & the world. Thus, the results cannot be used to comment on the effects of font on anyone who doesn't resemble the participants. Also, ~~no difference between~~ it cannot be concluded that serif font has no effect on reading speed until the study is carried out on a sample which more closely represents the general population. A relationship could exist but this specific sample doesn't allow it to be seen.

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

5a(i) The total number of new articles published per month on Wikipedia ~~increased from 0 in 2001 to~~ shows an overall increase from 0 in 2001 to about 20000 articles ~~per month~~ at the end of 2018. From 2001 ~~the~~ to 2007, the number of published articles increased ~~from about steadily~~ from 0 to 50,000, however since 2007, the number of published articles per month has been falling steadily (to 20000). Towards the end of 2002 there was a sharp peak in published articles to 38000, & and fluctuations were higher from 2006-2009 at about ± 10000 articles. However, other than these times, the fluctuations in articles have been relatively constant at ± 3000 articles. The no. of published articles show seasonality, with a peak each year in July to August. At other times of the year the number of published articles is relatively constant except for slight peaks in June, September and November.

5a(ii) The model forecasts 15196 articles published in November 2020, with lower and upper limits of ~~2020-11-30~~ -10365 and 40757 articles. ~~Because~~ However, a negative number of published articles is impossible.

The Holt-Winters model ~~using~~ uses a weighted average meaning that more weight is placed on recent data than old data. However, some weight is still placed on old data. In the raw data, as discussed before, the number of published articles increased from 2001-2007 before beginning to steadily decline since & then. This means that in using the data from 2001-2018, the old increasing trend would also have been included in making the forecast. This may be inappropriate since the recent raw data has been decreasing and is unlikely to increase in the future by much, and will probably level out. This is because as time goes on, there is more information on Wikipedia so less new info to publish. Hence, the forecast should have used only data from 2007 or 2009 onwards, to be more accurate based on the recent trend. If more recent data had been used, the forecast would probably be more level or & have a steeper decreasing trend.

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

Time use of an additive model is suitable since the seasonality has been constant over time ie. not requiring a ~~multiplicative~~ model. ~~However~~ Using the Holt-Winters model, as time goes on in the forecast, the 95% confidence interval gets larger. By November 2020 the interval is quite large (-10365, 40757) providing a large range of possible values. However, the forecast may have been affected by the greater fluctuations observed in 2007-2009. ~~The data from this time may have caused the~~ During this time, the great recession was happening so this could explain the large fluctuations. In 2009 the raw data was quite far from the fitted data so residuals would have been large at this time. These residuals may have affected the confidence intervals for the forecast & made them larger. Therefore, the model would be more suitable if it had been applied to only the more recent data rather than all eg. 2009 onwards.

However, the forecast is still only a prediction based on past trends and assumes that past trends and seasonality will persist. This may not be the case though as events like COVID-19 could have increased publishing due to more science research. ~~The reliability of~~

5b(i) I think that bar graphs have been used in the visualisation, with the length of the graph indicating the rankings of the book. The longer the bar, the higher the ranking ie. a ranking of 1 will show a bar double the length of a ranking of 10.

Colour coding has also been used to convey the ranking of the best selling fiction books. ~~Orange~~ when bars are orange this indicates that the book was ranked number 1. ~~Other bars are then etc.~~ Ranks after 1 are then ~~get~~ progressively lighter shades of blue, with lowest ranks (20) being light blue and higher ranks dark blue.

A timeline has been used on the horizontal axis to convey when the book had the rank denoted by the height/colour of its bar. If the

QUESTION NUMBER | Start each question on a new page. | ASSESSOR'S USE ONLY

are missing bars this means the book was not in the top 20 that week. The dotted lines represent a new year, and colour Therefore, the ~~their~~ ranking of a book is determined by bar height, while the bar's horizontal position ~~is determined~~ conveys the time at which it held that ranking.

5b(ii) The visualisation reveals that Fifty Shades of Grey held the number 1 ranking on the best sellers list for the most ~~same~~ number of consecutive weeks (about half the year from the second quarter of 2012 to 3rd quarter). Meanwhile, all the light blue cannot see held ~~the~~ a spot (any rank) on the best sellers list for the most consecutive weeks (more than a year, from the 4th quarter of 2014 to the second third of 2016. Most books are in the ~~top~~ best sellers list regularly for about 1-2 years. However, Gone girl was on the list regularly for 3 years. Furthermore, once a books hasn't been on the list for several months it often doesn't re-enter the list. This is with the exception of ~~Fifty Shades~~ the Fifty shades series - all 3 of which disappear from the list for up to a year but then re-enter it. Although, none of these books have been on the list since 2017. Arguably ~~thrones is the~~ all of the books on the visualisation have been made into movies except A game of thrones which is a TV show. A Game of Thrones is the only book that has been on the best sellers list every year since 2011, except 2018. This is probably because new seasons of the show come out every year, causing people to buy the book at these times and replace it in the best sellers list.

1

QUESTION NUMBER | Seen | Start each question on a new page. | ASSESSOR'S USE ONLY

*1a(iii) However, the outliers of the data first must be considered. The linear model was fitted to the data without the outliers being removed. The outlier at (2000,525) is unlikely to affect the ~~mean~~ model fitted. However, the outlier at (1900, 150) could have a considerable effect on the trend. This is because this point lies very far ~~about~~ (below) from ~~the~~ where majority of the data tends to lie. This could cause the trend line to be pulled lower (with a smaller gradient) than if the outlier was removed. If this point (and the ~~other~~ outlier were removed) it is highly likely that the gradient of the equation would increase above 0.2202. Therefore, it is possible that the gradient of a model with the ~~outer~~ outliers removed is very close to 0.250, ie. 250 wpm. Therefore, ~~therefore~~ the statement that 'for fiction books there are, on average, 250 words per page,' could be considered suitable. Provided that the outliers are infact very unusual and uncommon values, they could be removed to ~~provide~~ allow for a more accurately fitted linear model with a gradient close to 0.250.

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

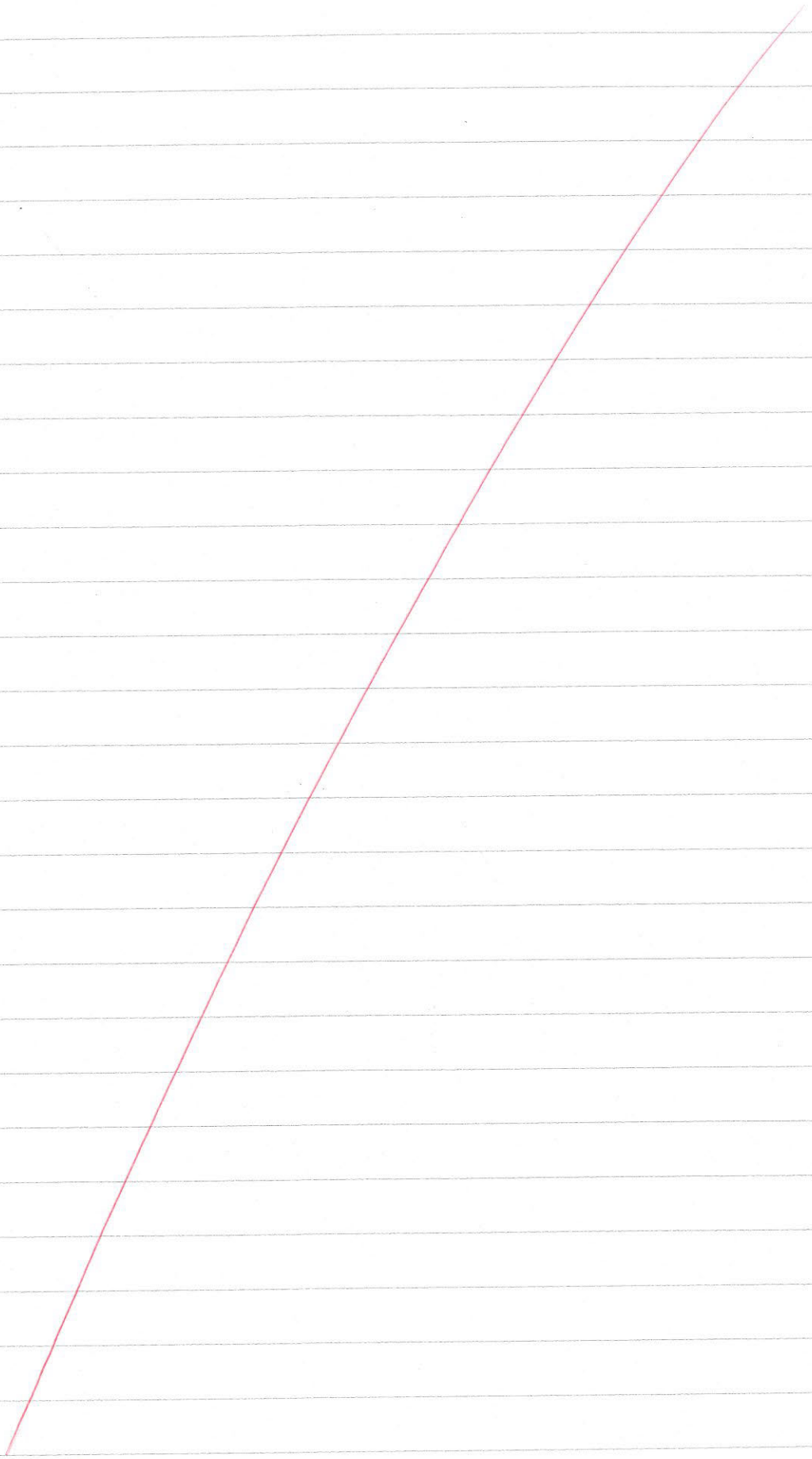
Start each question on a new page.

ASSESSOR'S USE ONLY

QUESTION NUMBER

Start each question on a new page.

ASSESSOR'S USE ONLY



93201A