(Answers calculated by RNC — *caveat emptor*. Thanks to MRFA for discussion and checking them.)

Q1 (2000 Paper 1) **Hidden assumptions — or, waffle you can skip in order to get to the answer.**

Before we answer the question, we must decide what kind of techniques to use. Should we use parametric or nonparametric techniques? **There's no necessarily 'right' answer.** We could summarize the theoretical arguments like this:

- The purist might say that the Beck Depression Inventory (BDI) is an **ordinal** rating scale; higher scores indicate 'more depression' somehow, but the differences between two ratings are not quantitatively meaningful (i.e. the difference between 0 and 10 is not necessarily the same as between 20 and 30) — like our example in Handout 1 of Army ranks (lieutenant → captain → major, etc.). This approach would mean that parametric techniques would not be applicable and we should use a nonparametric analysis.
- Alternatively, we might treat the BDI as an **interval** scale for the purposes of analysis, which would allow us to use statistics such as the mean and standard deviation, and appropriate parametric tests (assuming their other assumptions are met). There are several rationales for this, discussed for example by Velleman & Wilkinson (1993, 'Nominal, ordinal, interval, and ratio typologies are misleading for classifying statistical methodology', *The American Statistician* 47: 65–72), who argue that the 'meaning' of a scale is largely what you make of it — so if you're happy to speak about a '3-point change in a BDI score' then you should be happy to use parametric techniques here. As Francis Bacon (1620, *Novum Organon*) said, truth emerges more readily from error than from confusion. And while 'common' doesn't imply 'correct', it is certainly common to analyse rating scales with parametric techniques — the first paper on depression I looked at for this purpose analysed the Hamilton Depression Rating Scale using analysis of variance, which is a parametric technique (Mayberg *et al.* 2000 *Biological Psychiatry* 48: 830–843); the second I found did the same with BDI scores (Allen *et al.* 1998 *Psychological Science* 9: 397–401).

Being pragmatic, are there other barriers to using one or other technique?

- Part (b), comparing men and women before treatment, could be approached parametrically with an unpaired *t* test or non-parametrically with a Mann–Whitney *U* test, so no problem there.
- Part (c) could be approached parametrically with Pearson's *r* or non-parametrically with Spearman's r_s correlation, so no problem there.
- But part (a) asks which treatment is more effective — so we have to look at some difference between pre-treatment and post-treatment scores for each subject. If we take the differences (e.g. Post minus Pre scores) and analyse these, in whatever way, we have already made the assumption of an interval scale of measurement simply by calculating those differences, so we might as well use parametric techniques *unless their other assumptions are violated*. If we don't do this, the only information available is whether a subject improved or not. Since 15 / 15 Cogth patients improved, and 14 / 15 Couns patients improved, we're never going to find a significant difference with some form of categorical test (and a χ^2 test won't be valid since there will be expected values <5 and highly uneven across rows / columns, which violates the assumption of normality).

I'll illustrate both techniques below. Which will the examiners prefer? I've never been a 1B examiner, but if I were marking this, I'd accept either (particularly if some justification were given to show that you'd thought about the issue). You'll note that both actually give the same answers in terms of 'significant or not' judgements at conventional levels of α. And you may be influenced by the amount of work involved — ranking 30 scores is perhaps error-prone during an exam, whereas your calculator will do much of the work for the parametric tests.

The answer…

(P.T.O.)

Parametric version of Q1 (2000 Paper 1)

(a)

First, we calculate the difference between pre-therapy and post-therapy depression scores (Post – Pre), collapsing across (ignoring) sex. We find they are:

Cogth: {–10, –7, –11, –9, –13, –22, –18, –14, –12, –10, –11, –14, –9, –13, –11}

Couns: {–5, –4, +1, –6, –4, –8, –4, –7, –3, –9, –7, –8, –3, –8, –4}

The Beck Depression Inventory gives high scores to depressed people, and low scores to non-depressed people. So calculating the scores this way (Post – Pre), the better treatment will have the lower (more negative) difference score. The mean score in the Cogth group is –12.267 (n = 15, SD 3.770, variance 14.21); the mean score in the Couns group is –5.267 (n = 15, SD 2.658, variance 7.067). So the Cogth treatment appears to do a better job at reducing depression scores. Is this a significant difference? Let's run a two-sample unpaired t test. First, we run an F test to see if the difference between the variances is significant:

$$F_{n_1-1,n_2-1} = F_{14,14} = \frac{s_1^2}{s_2^2} = \frac{14.21}{7.067} = 2.01, NS$$

Since it wasn't, we can use the 'equal variances assumed' version of the two-sample unpaired t test, with the simpler formula since $n_1 = n_2$. So we can calculate t (with 28 df) as follows:

$$t_{n_1+n_2-2} = t_{28} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{(-12.27) - (-5.267)}{\sqrt{\frac{14.21}{15} + \frac{7.067}{15}}} = \frac{-7.003}{1.191} = -5.88, p < .01$$

Therefore, the Cogth group improved significantly more than the Couns group.

(b)

Before treatment, the men's scores were {42, 35, 32, 28, 26, 38, 33, 34, 29, 23} — a mean of 32 (n = 10, SD = 5.696, variance = 32.444). The women's scores were {20, 18, 17, 19, 21, 15, 22, 21, 27, 19, 19, 17, 18, 20, 23, 13, 24, 28, 24, 17} — a mean of 20.1 (n = 20, SD = 3.782, variance = 14.31). Let's run an unpaired t test again. First our homogeneity-of-variance check:

$$F_{9,19} = \frac{32.444}{14.31} = 2.267, NS$$

So we can use the 'equal variances assumed' version of the two-sample unpaired t test, but this time since the group ns are different, we must use the full formula:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{9 \times 32.444 + 19 \times 14.31}{28} = 20.14$$

$$t_{n_1+n_2-2} = t_{28} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{32 - 20.1}{\sqrt{\frac{20.14}{10} + \frac{20.14}{20}}} = \frac{11.9}{1.738} = 6.847, p < .01$$

So the men were significantly more depressed than the women before treatment.

(c)

To answer this question, we *correlate* the Pre (X) and Post (Y) scores in the Cogth group, and see if that correlation is significant. Our X–Y pairs are {20, 10}, {18, 11}, etc. If your calculator gives you r directly, fine. Otherwise, we'll use the long-winded formula for the covariance:

$$\text{cov}_{XY} = \frac{\sum(x - \bar{x})(y - \bar{y})}{n - 1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n - 1} = \frac{4779 - \frac{362 \times 178}{15}}{14} = 34.52$$

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{34.52}{7.52 \times 5.17} = 0.888$$

This is high, so likely to be significant; let's check that with the usual t test:

$$t_{n-2} = t_{13} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.888\sqrt{13}}{\sqrt{1-0.788}} = 6.954, p < .01$$

It is. There is a significant positive correlation (r = 0.888, p < .01) between depression before and after treatment in the cognitive therapy group.

(Beware: if you had found no correlation, you would not be able to say 'no' definitively to question (c) — remember, you would need to sketch a scatter plot, because r only measures *linear* relationships.)

Nonparametric version of Q1 (2000 Paper 1)

(a)

For reasons summarized above, treating the BDI as an ordinal scale is slightly half-baked here (in my opinion). The only reasonable way to approach it non-parametrically is to view the BDI as an interval scale for the purposes of calculating difference scores, but then to analyse these non-parametrically. First, we calculate the difference between pre-therapy and post-therapy depression scores (Post – Pre), collapsing across (ignoring) sex. We find they are:

Cogth: {–10, –7, –11, –9, –13, –22, –18, –14, –12, –10, –11, –14, –9, –13, –11}

Couns: {–5, –4, +1, –6, –4, –8, –4, –7, –3, –9, –7, –8, –3, –8, –4}

We then perform a Mann–Whitney *U* test. Both groups are the same size, so we'll arbitrarily call the Cogth group 'Group 1' (n = 15) and the Couns group 'Group 2' (n = 15). When we rank the scores from 1–30 we have the following ranks:

Group 1 (Cogth): 11.5, 20, 9, 14, 5.5, 1, 2, 3.5, 7, 11.5, 9, 3.5, 14, 5.5, 9

Group 2 (Couns): 23, 25.5, 30, 22, 25.5, 17, 25.5, 20, 28.5, 14, 20, 17, 28.5, 17, 25.5

So our sums of ranks are $R_1 = 126$, $R_2 = 339$.

We calculate

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 126 - \frac{15\times16}{2} = 6 \text{ and } U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 339 - \frac{15\times16}{2} = 219$$

so U = 6. (We also verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$ to check our arithmetic.) The critical value of $U_{15,15}$ is 65 for two-tailed $\alpha = 0.05$, so our U is significant. The Cogth group did significantly better than the Couns group.

(b)

Before treatment, the men's scores were {42, 35, 32, 28, 26, 38, 33, 34, 29, 23} (n = 15). The women's scores were {20, 18, 17, 19, 21, 15, 22, 21, 27, 19, 19, 17, 18, 20, 23, 13, 24, 28, 24, 17} (n = 20). We can compare these with a Mann–Whitney *U* test. We call the men's scores Group 1, because it's the smaller group. We rank all the scores (1–30). By group, they are:

Group 1 (M)	30, 28, 25, 22.5, 20, 29, 26, 27, 24, 16.5	*rank sum = 248*
Group 2 (F)	11.5, 6.5, 4, 9, 13.5, 2, 15, 13.5, 21, 9, 9, 4, 6.5, 11.5, 16.5, 1, 18.5, 22.5, 18.5, 4	*rank sum = 217*

We calculate

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2} = 248 - \frac{10\times11}{2} = 193 \text{ and } U_2 = R_2 - \frac{n_2(n_2+1)}{2} = 217 - \frac{20\times21}{2} = 7$$

so U = 7. (We also verify that $U_1 + U_2 = n_1 n_2$ and $R_1 + R_2 = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$ to check our arithmetic.) The critical value of $U_{10,20}$ is 56 for two-tailed $\alpha = 0.05$, so our U is significant. The men were significantly more depressed than the women before treatment.

(c)

To calculate Spearman's r_s, we **rank** the Pre and Post scores in the Cogth group. We have these raw scores:

Pre	20, 18, 17, 19, 21, 42, 35, 32, 15, 28, 22, 21, 26, 27, 19
Pre rank (X)	6, 3, 2, 4.5, 7.5, 15, 14, 13, 1, 12, 9, 7.5, 10, 11, 4.5
Post	10, 11, 6, 10, 8, 20, 17, 18, 3, 18, 11, 7, 17, 14, 8
Post rank (Y)	6.5, 8.5, 2, 6.5, 4.5, 15, 11.5, 13.5, 1, 13.5, 8.5, 3, 11.5, 10, 4.5

We correlate the ranks as X, Y pairs, e.g. {6, 6.5}, {3, 8.5}… If your calculator gives you r directly, fine. Otherwise, we'll use the formula for the covariance:

$$\text{cov}_{XY} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{1200.25 - \frac{120\times120}{15}}{14} = 17.16$$

$$r_s = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{17.16}{4.464\times4.452} = 0.863$$

For n = 15, the tables tell us that the critical value of $|r_s|$ is 0.689 at the two-tailed $\alpha = 0.01$ level, and our value is bigger than this, so our nonparametric correlation is significant at $p < .01$. So there is a significant positive correlation between depression before and after treatment in the cognitive therapy group.

Q2 (2003 Paper 1) *(a)*

We begin with a confidence interval question. Our sample has a mean of 614.8 and a standard deviation of 47.98, and n = 10. Since we've been told to assume a normal distribution and find the 95% CI, we can use our formula:

$$\text{Confidence intervals} = \bar{x} \pm \frac{s_X}{\sqrt{n}} t_{critical(n-1)df}$$

For 95% confidence intervals, we want critical values of t for α = 0.05 two-tailed, with n–1 = 9 *df*. From our tables, this critical value is 2.262. So we can plug that into our formula:

$$95\%\text{ confidence intervals} = 614.8 \pm \left(\frac{47.98}{\sqrt{10}} \times 2.262 \right) = 614.8 \pm 34.32$$

So there is a 95% probability that the true population mean lies within the range 580 to 649 (to 3 *sf*).

(b)

The second part of this question asks about a difference between groups. Regardless of all the irrelevant guff about the experimental design, the key point is that we have two independent groups of scores, n = 6 per group. So we'd like to run an two-sample unpaired t test. We can justify this since we've already been told that reaction times on this task are from a normally-distributed population (and we can see that they're not grossly non-normal). The first group has mean = 555.2, SD = 57.26, variance = 3279. The second group has mean 601.8, SD = 37.02, variance = 1370. First, we should check the variances are not significantly different with an F test:

$$F_{n_1-1,n_2-1} = F_{5,5} = \frac{s_1^2}{s_2^2} = \frac{3279}{1370} = 2.39, NS$$

Since the variances are not significantly different and $n_1 = n_2$, we can use our simple formula for the two-sample unpaired t test assuming equal variances:

$$t_{n_1+n_2-2} = t_{10} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{555.2 - 601.8}{\sqrt{\frac{3279}{6} + \frac{1370}{6}}} = \frac{-46.6}{27.84} = -1.674, NS$$

So the difference between groups was not significant (p > .1 two-tailed).

Q3 (2002 Paper 2) n / a — experimental design question.

See the end of this handout for advice on these questions.

Q4 (2002 Paper 1) A correlation and regression question…

(a) … estimate the rate of 'mental rotation' from the data.

The trick in the question is that the raw data you've given don't represent the rotation angles, since 'the transformation is thought to be carried out over the most direct route'. So first, we must calculate the corrected data:

original (°)	*0*	*45*	*90*	*135*	*180*	*225*	*270*	*315*
rotation (°)	**0**	**45**	**90**	**135**	**180**	**135**	**90**	**45**
RT (ms)	518	563	638	781	896	738	625	552
error (%)	0.87	0.84	1.47	2.44	4.20	2.29	1.33	0.53

We want to predict RT (Y) from the (corrected) rotation angle (X). We have a series of paired values, n = 8, and we'd like to find the regression equation

$$\hat{Y} = bX + a$$

We therefore want to find

$$b = \frac{\text{cov}_{XY}}{s_X^2} = r \frac{s_Y}{s_X}$$

and

$$a = \bar{y} - b\bar{x}$$

We can easily find $\bar{x} = 90^\circ$, $\bar{y} = 663.9$ ms, s_X = 58.92°, s_Y = 130.5 ms. Your calculator should also give you *r* directly; if not, calculate the covariance

$$\text{cov}_{XY} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{530{,}190 - \frac{720 \times 5311}{8}}{7} = 7457$$

We don't actually need to find *r* if we're doing it by hand, but a calculator will give us:

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{7457}{58.92 \times 130.5} = 0.970$$

So our regression coefficients are

$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{7457}{(58.92)^2} = r\frac{s_Y}{s_X} = 0.970\frac{130.5}{58.92} = 2.148 \text{ ms/degree}$$

$$a = \bar{y} - b\bar{x} = 663.9 - 2.148 \times 90 = 470.6 \text{ ms}$$

and our regression equation is

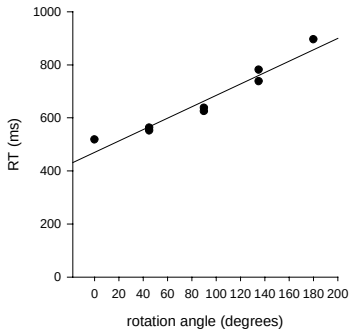
$$\hat{Y} = 2.148X + 470.6$$

I've tacked on the units of *a* and *b* above — you can easily work out what they must be, since you start with a number in degrees (*X*), and multiplying it by *b* gives you a number in ms (*Y*), and *a* must have the same units as *Y*. The reason for doing this is because these numbers are what you're actually after. The answer to 'what is the rate of mental rotation?' is 2.148 ms / degree (*b*)…

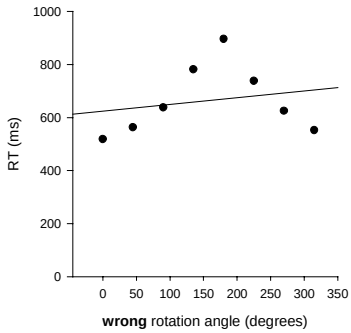
(a) What is your best estimate of the time occupied by the remaining components of the reaction time?

… and the remaining reaction time is 470.6 ms (*a*).

Although the question doesn't require it, you would be well advised to do a quick sketch of a scatterplot — for one thing, to see if there really is a linear relationship between *X* and *Y*:



A nice straight line. One benefit to doing a scatterplot in this question is that it might save you the embarrassment of failing to adjust the rotation angle. If you were in an exam-induced daze and didn't fix the angles, your scatterplot would look like this:



… which certainly isn't a linear relationship, and might make you notice that something was amiss.

(c) Is there a significant relationship between reaction time and error rate?

Here we go again — but this time, the question only asks 'is there' a relationship, not 'what is' the relationship. So we only need to calculate a correlation and test its significance. This time let's call RT *X* and error rate *Y* (or as you see fit). Either get *r* from your calculator or calculate

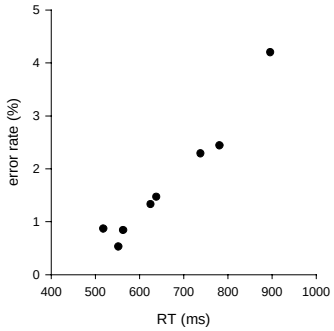
$$\text{cov}_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{10{,}344.11 - \frac{5311 \times 13.97}{8}}{7} = 152.83$$

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{152.83}{130.5 \times 1.202} = 0.974$$

Such a high correlation is certain to be significant with 8 observations. To test this formally, we calculate a *t* score:

$$t_{n-2} = t_6 = \frac{0.974\sqrt{6}}{\sqrt{1-0.9487}} = 10.53,\ p < .01$$

(since the critical value for *t* with 6 *df* is 3.707 for a two-tailed α = 0.01). So there is a significant relationship between RT and error rate. The scatterplot **(not required but a sketch may be helpful)** is shown below.



But please note: if there had been no significant correlation, you would not have been able to say there was no relationship, just no linear relationship. This is where scatterplots help (is there a nonlinear relationship?), so you're always advised to sketch one, however rough it is.

*This final part is **not** part of the question… but you might be wondering whether there's just a relationship between RT and error rate because RT is related to the rotation angle, and errors are related to the rotation angle. We'd work out the correlation between RT (call it X) and error rate (call it Y), 'partialling out' the effects of rotation angle (call it Z). First we need to work out the one correlation we haven't worked out yet — between rotation angle (corrected so all the angles are ≤180°, of course) and error rate. It's 0.909. Then we can calculate*

$$r_{xy.z} = \frac{r_{xy} - r_{xz}r_{yz}}{\sqrt{(1-r_{xz}^2)(1-r_{yz}^2)}} = \frac{0.974 - 0.970 \times 0.909}{\sqrt{(1-0.970^2)(1-0.909^2)}} = \frac{0.09227}{0.1013} = 0.911$$

So even accounting for the relationships between RT and rotation angle, and between error rate and rotation angle, there's a pretty strong relationship between RT and error rate (though it's less than the correlation we first worked out of 0.974).

Q5 (2001 Paper 2) n / a — experimental design question.
See the end of this handout for advice on these questions.

Q6 (2001 Paper 1) *(a) Recognition v recall*

There are some hidden ambiguities in this question.

First off, we want to test if subjects differed for recognition and recall scores. This calls for either a paired *t* test (parametric) or a Wilcoxon matched-pairs signed-rank test (non-parametric). If the assumptions of the *t* test are met (i.e. the differences between each pair of data come from a normally-distributed population), the *t* test has more power.

The difference scores (recall minus recognition) are 7, –11, –10, –7, –10, 0, 0, –18, –1, 3.
Placed in order, they are –18, –11, –10, –10, –7, –1, 0, 0, 3, 7.

Parametric or nonparametric? A quick glance (or histogram, or stem-and-leaf plot) suggests that this isn't the best normal distribution in the world, so you may favour the Wilcoxon test. But note this: later, the question wants you not only to correlate these variables (which could be accomplished either with Pearson's *r* or Spearman's r_s) but to perform a regression, for which you only know parametric techniques (which assume normality of both the recall and recognition scores and that the difference between two normal random variables should also be normally distributed).… so you could probably argue the case either way.

If you had run a one-sample *t* test, you'd have got this: the difference scores have a mean of –4.7 and an SD of 7.689; $n = 10$. So

$$t_{n-1} = t_9 = \frac{\bar{x} - \mu}{\frac{s_X}{\sqrt{n}}} = \frac{-4.7 - 0}{\frac{7.689}{\sqrt{10}}} = \frac{-4.7}{2.431} = -1.933$$

From tables, you'd have found that $p < .05$ for a one-tailed test but $.05 < p < .1$ for a two-tailed test.

If you'd used the Wilcoxon test, there are 8 non-zero difference scores and you'd have found $T_8 = 5.5$, for which $p < .05$ one-tailed but $.05 < p < .1$ two-tailed.

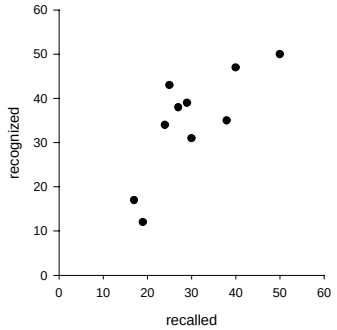
So both approaches give the same answer (which, if you actually chose to run them both, should reassure you that they're giving valid answers).

But what of the original question — 'determine whether recognition performance is better than recall performance'? This phrasing suggests a one-tailed test, and recognition performance did give the higher scores, so you may say 'significant, $p < .05$ one-tailed'. Alternatively, you may decide that a real researcher would not want to ignore a difference in the opposite direction, and say 'not significant, $p > .05$ two-tailed'. Since you wouldn't want to be accused either of misinterpreting the question or answering a slightly daft scientific question, you could just say 'one-tailed $p < .05$ but two-tailed $.05 < p < .1$', showing that you know what you're talking about, and then choose and defend a choice of a one- or a two-tailed test. It's your understanding of what's going on that counts, rather than any particular choice you justify sensibly.

(b) scatterplot, correlation, regression

'Construct a scatter plot of the number of pictures recognised against the number recalled.'

You could assign either to the *X* and *Y* axes. Looking ahead, we're going to be asked to predict recognition from recall, so let's call recall *X* and recognition *Y* so we're on familiar ground, predicting *Y* from *X*.



'Did those subjects who recalled more pictures also recognise more of them?'

That's asking 'was there a significant correlation between recall and recognition?' Either calculate *r* with your calculator, or work it out like this. Again, I'll call recall *X* and recognition *Y*:

$$\text{cov}_{XY} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1} = \frac{\sum xy - \frac{\sum x \sum y}{n}}{n-1} = \frac{11{,}205 - \frac{299 \times 346}{10}}{9} = 95.51$$

$$r_{XY} = \frac{\text{cov}_{XY}}{s_X s_Y} = \frac{95.51}{10.14 \times 12.14} = 0.776$$

Is this significant? We work out a *t* score:

$$t_{n-2} = t_8 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.776\sqrt{8}}{\sqrt{1-0.602}} = 3.479,\ p < .01$$

So the answer's yes; there is a significant positive correlation (subjects who recognized more also recalled more).

'Plot on your graph the line that best predicts recognition performance from recall performance.'

We need to calculate the regression equation

$$\hat{Y} = bX + a$$

We therefore want to find

$$b = \frac{\text{cov}_{XY}}{s_X^2} = r\frac{s_Y}{s_X}$$

and

$$a = \bar{y} - b\bar{x}$$

So we have

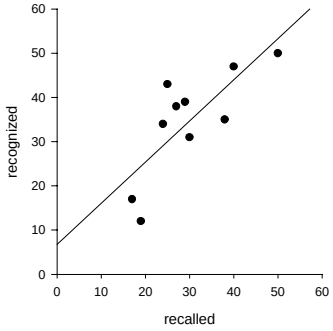
$$b = \frac{\text{cov}_{XY}}{s_X^2} = \frac{95.51}{(10.14)^2} = r\frac{s_Y}{s_X} = 0.776\frac{12.14}{10.14} = 0.929$$

$$a = \bar{y} - b\bar{x} = 34.6 - 0.929 \times 29.9 = 6.823$$

So our regression equation is

$$\hat{Y} = 0.929X + 6.823$$

We can plot this on our scatterplot, and we're done.



Q7 (2000 Paper 2) n / a — experimental design question.
See the end of this handout for advice on these questions.

Q8 (2003 Paper 2) n / a — experimental design question.
See the end of this handout for advice on these questions.

About the experimental design questions

There's no 'right' answer to an experimental design question. Good experimental design requires that you understand the question well — if you don't know what retroactive and proactive interference are, for example, you'll have difficulty with Q3(c). But *in general*, what things should you think about when designing experiments?

One vital thing is to establish **what you'll measure.** *What numbers will you actually write down?* For example, for question Q3(a), will you measure baby looking times? Baby approach distances? Proportion of occasions on which the baby approaches? Another example: Q8 (part 2) asks about investigating whether aspirin actively affects the operation of the active mechanism in the cochlea. Will you measure detection thresholds at different frequencies? Will you measure frequency selectivity (and how)? Will you measure otoacoustic emissions? To choose, you need to know what aspects of hearing actually depend on the active mechanism of the cochlea — obviously, not all aspects of hearing do.

You also need to determine your **subjects.** Will you use people? Owls? Rats? Dissected cochleas? If you use humans, who? Psychology undergraduates? People recruited from newspaper adverts? Will you impose restrictions on the age or sex of your subjects? Will you exclude them if they have a history of mental illness, head injury, or ototoxic antibiotic use? Sometimes your experimental technique influences this: you can't give PET scans to young women (potential egg damage from radiation), you can't put people with implanted magnetic metal into an MRI scanner (it accelerates hard and tends to kill), and you might be careful before giving a drug that makes people unhappy to those with a history of severe depression.

Will you use a **correlative** or a **causal** technique? In our aspirin example, will you seek out people who use lots of aspirin and compare them to people who don't? If so, do you need to match these groups somehow? What might confound your interpretation of any differences? Or will you use a causal technique, in which you give aspirin to subjects in some fashion and compare them to people who didn't receive aspirin?

Many psychological experiments are based on simple intervention studies, where treatments are controlled by the experimenter. If done properly, these allow inferences to be made about whether the treatment *caused* an effect. They may use between-subjects (between-groups) or within-subjects designs

- A simple *between-groups* design: assign subjects (at random) to groups. Treat each group differently, before giving them all the same test. If the test results for the groups are different, this is evidence that the treatment influences performance.
- A simple *within-subjects* design: test the same subjects repeatedly after (or during) different treatments, counterbalancing for the order in which you give the treatments and so on. If the results differ across treatment conditions, this is evidence that the treatment influences performance.
- Sometimes we have to use more complex experimental designs. For example, when more than one treatment is given, they can all be given in a between-subjects fashion, or all in a within-subjects fashion, or some between- and some within-subjects.
- Sometimes we have to consider differences between groups that are not assigned by the experimenter — such as gender, age, IQ, prior illness, prior drug use. The exact interpretation of differences between groups in such an experiment may be more complicated, as the effect of one variable may be **confounded** by another. For example, if you give all the male subjects treatment A, and all the female subjects treatment B, you won't be able to tell whether a difference between the groups is due to the treatment difference or the sex difference — these two variables are confounded. In general, **random assignment** of subjects to groups is a good way to get around this problem.

So will you use a **between-subjects** or a **within-subjects** design? In our aspirin example, will you test subjects on aspirin and the same subjects off aspirin? Or will you give some subjects aspirin and some not?

- If you use a within-subjects design, will you test 'off' aspirin and then 'on'? Or 'on' then 'off'? Will there be effects of practice on the task that affect the interpretation in this case? Will the drug have permanent or long-lasting effects? Should you randomize or counterbalance the order (so some subjects get aspirin first and some get placebo first)?
- When you're not giving the drug, will you give nothing, or a placebo? If you use a pla-

cebo, will the experimenter be 'blind' as to the condition the subject is in? The importance of the **experimenter's awareness,** or lack of it, applies more generally whenever there is the possibility that the experimenter's expectations may influence the subject, or influence the recording of the data, consciously or unconsciously.

- If you use a between-subjects design, how will you assign subjects to groups?

Sometimes the question asks about how you will **analyse** the data you collect. The design of your experiment partly determines the analytical technique. Do you have **quantitative** or **categorical** data? The use of a between-subjects or a within-subjects design influences the 'relatedness' of your data, and may determine whether you will use **related** (e.g. paired) or **unrelated** (unpaired) statistical tests. What will your **null hypothesis** be? There may be some things you can't specify in advance (e.g. you may prefer to use parametric tests like the *t* test, but you don't always know whether their assumptions will be met until you collect the data; if their assumptions are violated, you may need to use nonparametric tests instead).

Good designs are **simple,** and answer the question clearly. If you find an effect in your experiment, will the **interpretation** be simple?

Sometimes you may need to design a **series of experiments** rather than a single experiment. Think about what each experiment should aim to establish; keep each experiment as simple as possible. Sometimes the most sensible choice of the next experiment depends on the results of previous experiments; you can do no more than anticipate likely outcomes or lines of investigation if you are outlining a proposed series of experiments.

Good designs are also **practical.** If your design calls for the use of a zero-gravity cell culture environment, it may be impractical or expensive — but if the question is important (will it save lives? improve the lot of millions?), maybe it's worth it. On the other hand, if you could do the same experiment with a set of headphones and a signal generator, that's probably preferable. If your design involves inducing permanent hearing damage in volunteer humans, it's highly questionable **ethically.** When using animals, experimenters always seek to *refine* experiments to minimize suffering and distress, *reduce* the number of animals used, and *replace* live animals with alternatives when possible.

Finally, your design may be excellent, or it may just be the best thing you can think of during the exam. If you spot **problems** or flaws in your design, discuss them. You're not expected to design something perfect, but if you know there are problems, talk about them.

Many of these issues were also discussed in the *Background Knowledge* handout that you received at the start of the course (available with all the other statistics handouts, slides, etc., at **www.pobox.com/~rudolf/psychology**).

Glossary of jargon

You may encounter the following terms in relation to experimental design. This glossary is here to help you to understand descriptions of experiments; you don't have to use the jargon for your own answers (though you can if you want).

- **Between-groups design.** Same as *between-subjects design.*
- **Between-subjects design.** A design in which individual subjects are each given only one treatment; different treatments are given to different groups of subjects. For example, we might assign subjects to group A or group B; each individual has their performance on a task measured after being given either sugar (for group A), or amphetamine (for group B). See also *within-subjects design.*
- **Blind.** Unaware. 'Blind to' means 'unaware of'. For example, in a *single-blind* drug study, the subject is unaware of whether he has taken an active drug or a *placebo*. If the subject is not blind, his expectations of the treatment's effects may influence the results. For example, if the subject is a depressed patient in a study examining the effects of a new antidepressant, he may expect good things of the drug, and therefore feel good about the fact that he's receiving it, and feel less depressed. See also *placebo, double-blind*.
- **Clever Hans effect.** The fact that subtle and unintentional cueing by an experimenter, which reflects the experimenter's own expectations, may influence subjects. Wilhelm von Osten was a retired German schoolmaster who attempted to teach his Russian stallion Hans arithmetic. Von Osten would show Hans numbers on cards; the horse would tap out the number, or the answer to simple mathematical problems, with its hoof. A group of observ-

ers including a zoologist, a vet, and a politician were convinced; they could detect no fraud or cueing on the part of von Osten. Finally, Oskar Pfungst conducted a very thorough series of experiments in 1907 to investigate Hans's performance [Pfungst, O. (1907) *Das Pferd des Herrn von Osten (Der Kluge Hans). Ein Beitrag zur experimentellen Tier- und Menschen-Psychologie*; Leipzig: J.A. Barth]. First, he established that if van Osten did not know the number or answer, Hans did not succeed. Second, he showed that if the horse could not see van Osten, it also failed to get the right answer. Pfungst then turned his attention to van Osten, and noticed that he made almost imperceptible alterations in posture when interrogating the horse. Van Osten inclined his head as the horse began to tap the ground, and straightened slightly, lifting his eyebrows and flaring his nostrils slightly, when the horse approached the correct answer. The horse had learned to respond on the basis of these cues. Pfungst then stood in front of the horse himself, remained silent, showed no cards, and yet made Hans tap his hoof and cease using slight head movements. Pfungst went on to perform experiments in which he played Hans's role; he showed that over 90% of human subjects provided subtle bodily cues as the correct answer was approached (attributed to 'tension release'), just as von Osten had. See also *double-blind*.

- **Confound.** Two variables are confounded when their effects are impossible to distinguish. Suppose we want to establish whether a drug (call it treatment A) influences performance on a particular task, compared with placebo (call it treatment B). If you give all the male subjects treatment A, and all the female subjects treatment B, you can't tell whether a difference between the groups is due to the treatment difference or the sex difference — these two variables are confounded. In general, random assignment of subjects to groups is a good way to get around this problem (see also *randomization*). Common confounding factors worth thinking about are time (see *order effects*) and who collects the data.
- **Control.** OED: 'A standard of comparison for checking inferences drawn from an experiment; specifically a patient, specimen, etc., similar to the one(s) being investigated but not subjected to the same treatment.' If two groups are compared and one receives some critical treatment but the other does not, we refer to the latter as the *control group*. (See also *placebo*, *sham*.)
- **Counterbalancing.** A method of avoiding *confounding* among variables. Suppose subjects are tested on both an auditory reaction time task and a visual reaction time task. If all the subjects are tested on the auditory task first, the task order (first versus second) is confounded with the task type (auditory versus visual), and *order effects* such as a *practice effect* may account for any observed differences. The experiment should have been designed better: task order and task type should have been counterbalanced, such that half the subjects were given the auditory task first and half were given the visual task first. When there are several conditions, a *Latin square* is often use to design the counterbalancing.
- **Dependent variable.** A variable that you measure, but do not control. Compare *independent variable*.
- **Double-blind.** Where both the experimenter and the subject are unaware of the treatment that the subject receives. Otherwise, either the experimenter's expectations, or the subject's, or both, may influence the results. For example, if Alice is interviewing patients about their mood as part of a study looking at antidepressant effects, and is aware that the patient has been on a new drug of which Alice has a high opinion, she may expect the patient to be happier, therefore be and appear happier herself, and therefore influence the answers. See *clever Hans effect*. The 'gold standard' for a medical drug trial is a *double-blind placebo-controlled* study (see also *placebo*, *control*).
- **External validity** (also known as **generality** or **applicability**). The degree to which your experimental results can be applied to other populations and settings. If you examine reaction times in Cambridge 1B psychology undergraduates, to what extent can your results be generalized to Cambridge undergraduates? To UK undergraduates? To undergraduates in general? To people in general? Compare *internal validity*.
- **Factorial design.** When an experimenter is interested in the effects of two or more treatments, it is common to analyse them in a factorial design. Suppose we are interested in the effects of nicotine on psychomotor performance. We might be interested in the effects of both (1) nicotine dose, and (2) the difficulty of the task. So one variable ('factor') is drug dose; the other is task difficulty. Suppose there are three doses (none, low, high) and two levels of task difficulty (easy, hard). In a factorial design, we test every combination of the factors (i.e. none / easy, none / hard, low / easy, low / hard, high / easy, high / hard). Factorial designs can have more than two factors. We have *not* covered the statistical techniques required to analyse factorial designs, but they are widely used in research.
- **Independent variable.** A variable that you control or manipulate (e.g. drug versus placebo). Compare *dependent variable*.

- **Internal validity.** To what degree are you justified in drawing conclusions from your data? Basically, was your experiment any good? If you have overlooked a *confound*, you may be unable to interpret your data in the way you had hoped, and you do not have good internal validity.
- **Interpretation bias.** When the interpretation of evidence is influenced, sometimes inappropriately, by prior beliefs. *Confirmation bias* — people's tendency to notice and remember evidence that confirms their beliefs or decisions, and to ignore, dismiss, or forget evidence that is discrepant. *Rescue bias* — discounting data by finding selective faults in the experiment. *Auxiliary hypothesis bias* — introducing *ad hoc* modifications to imply that an unanticipated finding would have been otherwise had the experimental conditions been different. *Mechanism bias* — being less sceptical when underlying science furnishes credibility for the data. *"Time will tell" bias* — the phenomenon that different scientists need different amounts of confirmatory evidence. *Orientation bias* — the possibility that the hypothesis itself introduces prejudices and errors and becomes a determinate of experimental outcomes. [See Kaptchuk TJ, 2003, BMJ 326: 1453–5.]
- **Latin square.** (You don't need to know this!) A way to *counterbalance* the order in which subjects are tested in different conditions. A Latin square is a *n* by *n* grid in which each of *n* symbols appears exactly once in each row and once in each column. This can be used to allocate subjects to different treatment orders. If there are four treatments (A, B, C, D) and each subject must be tested in each condition once, a good Latin square is ABCD, CADB, BDAC, DCBA. If you randomly assign subjects to these four treatment orders, you will have successfully counterbalanced for treatment order. Each treatment immediately precedes and follows the other conditions once, known as a digram-balanced Latin square. (A much less good Latin square is ABCD, BCDA, CDAB, DABC, known as cyclic.)
- **Matching.** One way to avoid *confounds* between variables. Suppose we want to compare the performance of two groups of people on a reaction-time task; one group will receive drug and the other will receive placebo. We want our groups *not* to differ in any variable except drug v. placebo — otherwise our treatment and that variable will be confounded. Matching would involve deliberately measuring all sorts of things that we think might be relevant (e.g. IQ, age, sex, reaction time on a different task…) and assigning subjects to groups so that both groups have a similar distribution of sex, age, IQ, and so on. See also *randomization*, which is the other very important method to use in this situation; randomization deals with all the other variables you haven't thought about.
- **Order effect.** A concern of *within-subjects* designs, in which subjects are each tested several times: the order in which you test subjects may be an important factor. This may be due to *practice* (performance getting better with time), or other factors such as fatigue or boredom (getting worse with time), lingering drug effects (e.g. drug not yet fully gone from the body from a previous occasion; tolerance develops to drug with time), etc.
- **Orientation bias.** When prior expectations influence the collection of data. For example, psychology graduate students, when informed that rats were specially bred for maze brightness, found that these rats outperformed those bred for maze dullness, despite both groups really being standard laboratory rats assigned at random. [See Kaptchuk TJ, 2003, BMJ 326: 1453–5.]
- **Placebo.** Literally, 'I shall please'; a pill or procedure prescribed by a doctor for the psychological benefit of obtaining a prescription, rather than any physiological effect. Often a sugar pill. A *placebo effect* is an effect caused by a placebo. In research design, to establish the effect of a drug, one must compare it with something similar in all respects except for the drug itself. Therefore it's not wise to give one group of people the drug and the other (control) group nothing; the control group should be given a placebo. If the drug is in pill form, the placebo should be an inactive pill, labelled identically; if the drug is an injection, the placebo should be some inactive substance that is also injected (e.g. saline solution, or whatever liquid 'vehicle' the drug is dissolved in).
- **Practice effect.** If subjects are tested repeatedly, they may get better as a result of practice. An example of an *order effect.* Suppose you want to measure the effect of amphetamine on performance of the computer game Tetris, using a *within-subjects design*. You could give your subjects a *placebo* and test them; you could then give them amphetamine and retest them. Suppose they're better the second time: is this due to the drug or to the effects of practice? You can't tell: the two are *confounded.* You should have *counterbalanced* the order assignment.
- **Publication bias.** The tendency of referees and journal editors (and sometimes the scientists concerned) to publish studies (or not) based on the direction or strength of the study's findings. For example, journals like to report 'significant' effects, because they seem more interesting, meaning that studies failing to find that a treatment has an effect may be less

likely to be published. As a consequence, too high a proportion of what you read in journals would suggest the treatment *does* have an effect.

- **Randomization.** Random assignment of subjects to groups and / or treatment conditions is an important way to avoid inadvertent *confounds* (q.v.). Suppose we want to compare the performance of two groups of people on a reaction-time task. One group will receive a drug and the other will receive a placebo. We want our groups *not* to differ in any factor except that which we're manipulation — otherwise our treatment and that variable will be confounded. We might attempt to *match* groups (see *matching*) for relevant variables. But we probably can't explicitly match groups on every variable that might potentially be a confound; eventually we need a mechanism to decide which group a subject goes in, and that method should be random assignment. So in our example, if we have plenty of subjects, we could just randomly assign them to the drug group or the placebo group. Or we could match them a bit better by ranking them in order of reaction time performance and, working along from the best to the worst, take pairs of subjects (from the best pair to the worst pair), and from each pair assign one to the drug group and one to the placebo group at random. Random assignment takes care of all the factors you haven't thought of — for example, if your subjects are all going to do an IQ test in your suite of testing rooms, you should seat them randomly, in case one room's hotter than the others, or nearer the builders' radio outside, or whatever.
- **Sham.** Similar to *placebo*; the term is often used to refer to practical (e.g. surgical) procedures. If one group of rats receives amygdala lesions, the appropriate control group is probably not a set of unoperated rats, but a set of rats who have received 'sham' surgery — surgery identical except for the omission of the toxin that destroys the amygdala, say.
- **Within-subjects design.** A design in which individual subjects are each given more than one treatment, at different times. For example, each individual has their performance on a task measured after being given sugar, and on a separate occasion after being given amphetamine. These designs are often statistically powerful (they need fewer subjects to detect effects of the treatment), since differences between subjects' ability to perform the task don't contribute to the measurement error. Problems with these designs include *practice* and *order* effects; attention must be paid to proper *counterbalancing* of the order in which subjects are tested in different conditions.