Pre-test and Post-test scores

The researcher is usually interested in seeing how well students perform under a variety of teaching methods, and they have pretest scores on the students prior to the learning phase.

How should I analyze this type of data?

1. Should I enter the pretest score as a covariate into the model?

2. Should I just use a difference or ‘improvement’ score as posttest-pretest?

3. Should I fit a split-plot analysis where a student is nested in a method and then observed at two time points (pre and post)?
Simulation: Comparing teaching techniques by measuring ‘ability’ with a Pretest and Posttest under two teaching conditions (methods)

In my simulation, I assumed we could measure Pretest score ($x$-covariate) without measurement error.

The simulated true model in words:

A Posttest score equals the Pretest score plus the ‘mean’ gain from the method plus an error term.

The simulated true model in statistical notation:

$$Posttest_i = Pretest_i + \beta_0 + \beta_D D_{\text{group}_i} + \epsilon_i$$

$D_{\text{group}}$ is dummy variable for trt group (0=control).

$\beta_0$ represents the ‘mean’ benefit of being taught under the control method (gain over Pretest).

$\beta_D$ represents the mean difference between the treatment and control benefit.

$\epsilon_i \sim iid \, N(0, \sigma^2)$

$\epsilon$ includes variability due to both Posttest measurement error (we’re trying to measure ‘ability’, a latent variable) and Posttest variability among individuals with the same Pretest score and treatment group.
Can I perceive this as a t-test on the differences? i.e. Do the two groups have different improvements?

The true simulated model re-ordered:

\[ Posttest_i - Pretest_i = \beta_0 + \beta_D D_{\text{group}_i} + \epsilon_i \]

which is the same as...

\[ \text{Diff}_i = \mu + \alpha_{\text{group}_i} + \epsilon_i \]

THAT’S A T-TEST ON THE DIFFERENCES!!
Simulated values
n=100 with 50 in each treatment group.
Pretest scores randomly generated from $N(60, 10^2)$.
Mean benefit from being in control group = 12.
Mean benefit from being in treatment group = 20.
And $\sigma = 5.65$

Pretest score (baseline information) disappears in the differences.
Plot of the differences:

![Plot of the differences](image)

> `t.test((posttest-pretest)~trt,var.equal=TRUE)`

Two Sample t-test

data:  (posttest - pretest) by trt
t = -6.5223, df = 98, p-value = 3.042e-09
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  -10.189924   -5.435674
sample estimates:
mean in group 1 mean in group 2
  11.65641       19.46921
Can I perceive this as an ANCOVA model?
The true simulated model re-ordered:

\[ Posttest_i = \beta_0 + 1 \times Pretest_i + \beta_D D_{\text{group}_i} + \epsilon_i \]

(baseline covariate)

Looks like an ANCOVA with parallel lines among groups (no interaction term between pretest and group) and a covariate slope coefficient forced to be a 1.

Fitted ANCOVA below (no restriction on covariate coefficient):

\[ \text{lm.out.add=lm(posttest~trt+pretest)} \]
\[ \text{summary(lm.out.add)} \]

ANCOVA model output:

\[
\text{Coefficients:}
\begin{array}{cccccc}
\text{Estimate} & \text{Std. Error} & \text{t value} & \text{Pr(>|t|)} \\
(Intercept) & 7.42934 & 3.89922 & 1.905 & 0.0597 . \\
trt2 & 7.89301 & 1.19862 & 6.585 & 2.34e-09 *** \\
pretest & 1.07070 & 0.06366 & 16.819 & < 2e-16 *** \\
\end{array}
\]

NOTE: \text{pretest} coefficient not significantly different than 1, \( t_{obs} = 1.11, p\text{value}=0.2697 \):
Any difference in ANCOVA and $t$-test results?
1000 simulations under $H_A$: Record $t$-test for treatment

$t$-test $p$-values vs. ANCOVA $p$-values (treatment effect):

- $\log_{10}(pvalue)$: larger $\Rightarrow$ more significant

80.5% had smaller $p$-value in ANCOVA

The point (8,5) is used for $p$-values ($10^{-8}, 10^{-5}$)

OLS estimates and standard errors in ANCOVA.

In $t$-test, difference in improvement divided by $\sqrt{\frac{\sigma^2}{50} + \frac{\sigma^2}{50}}$

Not exactly the same results, but strongly correlated.
Estimated Treatment effect:

Another ANCOVA output example (coefficient NOT forced to be 1):

```r
> lm.out.add=lm(posttest~trt+pretest)
```

Coefficients:

|                | Estimate | Std. Error | t value | Pr(>|t|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | 2.2208   | 36.8012    | 0.060   | 0.952    |
| trt2           | 8.3796   | 1.2446     | 6.733   | 1.18e-09 *** |
| pretest        | 1.1668   | 0.6132     | 1.903   | 0.060    |

Residual standard error: 6.184 on 97 degrees of freedom
Multiple R-squared: 0.351, Adjusted R-squared: 0.3376
F-statistic: 26.23 on 2 and 97 DF, p-value: 7.835e-10

NOTE: pretest coefficient not significantly different than 1, \( t_{obs} = 0.332, pvalue=0.7406 \):
What about Type I error rate in ANCOVA?

For how the data were simulated, this a model mis-specification. Is the ANCOVA declaring significance too often?

1000 simulations under $H_0$ with no teaching method effect.

<table>
<thead>
<tr>
<th>Proportion of times effect declared significant using $p = 0.05$ cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANCOVA</td>
</tr>
<tr>
<td>0.049</td>
</tr>
</tbody>
</table>

Looks OK, but I think I would personally still recommend $t$-test on differences or split-plot analysis (next) as these models test for the effect of interest in a way that is consistent with how I perceived the data were generated.
The ANCOVA model (parallel lines):

\[ Posttest = \beta_0 + \beta_1 Pretest + \beta_D D_{group} + \epsilon \]

As we saw before, if we force \( \beta_1 = 1 \) in this ANCOVA, then this is the exact same testing for treatment (i.e. pvalue) as a difference in improvement between the two groups.

What does forcing \( \beta_1 = 1 \) imply in the ANCOVA?
A 1 point increase in \( Pretest \) score is associated with a 1 point increase in \( Posttest \) score (on average).

Seems like setting \( \beta_1 = 1 \) is at least reasonable.

What does excluding interaction \( Pretest \times D_{group} \) imply?
A 1 point increase in \( Pretest \) score must have the same impact on \( Posttest \) score for both groups.

Overall, as long as you do not have reason to believe that the slope for \( Pretest \) should differ between groups and a 1 point increase in \( Pretest \) coincides with a 1 point increase in \( Posttest \), then the \( t \)-test on differences seems most straight-forward when testing for treatment effect.

But if these assumptions are questionable, we can use ANCOVA with interaction to give a more flexible model.
Can I perceive this as a split-plot model?

Student is nested in a treatment group, and we observe the student at two time points.

Example:

If students are randomly assigned to treatment group, we expect no difference between the groups at the first time point. If there is a teaching method effect, we expect to see an interaction between treatment group and time.

In other words, we’re looking to see if the difference between the pre and post scores differs across treatment groups.
R output: t-test vs. split-plot

# t-test on the differences:
> t.test((posttest-pretest)~trt,var.equal=TRUE)

Two Sample t-test

data: (posttest - pretest) by trt
t = -5.0543, df = 98, p-value = 2.008e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -10.114986 -4.411446
sample estimates:
mean in group 1 mean in group 2
 12.42693  19.69015

## Split-plot:
> subject=1:100
> time=as.factor(c(rep("pre",n),rep("post",n)))
> time=factor(time,levels=c("pre","post"))
> sp.data=data.frame(trt.group=as.factor(rep(trt.grouping,2)),
>                      subject=as.factor(rep(subject,1)),score=c(pretest,posttest),time)
> aov.out=aov(score~trt.group*time+Error(subject),data=sp.data)
> summary(aov.out)

Error: subject
     Df Sum Sq  Mean Sq  F value    Pr(>F)
trt.group    1 1398.6 1398.6    5.543 0.020550  *
Residuals   98 24725.5   252.3

---

Error: Within
     Df Sum Sq  Mean Sq    F value    Pr(>F)
  time     1 12893.8 12893.8 499.494862 < 2.2e-16  ***
trt.group:time 1  659.4   659.4  25.546717 2.008e-06  ***
Residuals 98 2529.7   25.8

---
1000 simulations: split-plot p-values (interaction) vs. t-test p-values (on differences)

These are doing the same test.
But, as always, listen to your client and be understanding of the things that are commonly used and accepted (and even expected) within each discipline.

In this case, the t-test on the differences and the ANCOVA give results that are fairly similar. So, if they’re most comfortable with using a pre-test score as baseline covariate, then that’s how I would proceed. And in the education discipline, I believe the ANCOVA model is common.

NOTE: The covariate used in the ANCOVA is commonly centered in order to give a nice interpretation of the intercept. When Pretest is centered, $\beta_0$ is the mean Posttest score for the baseline group at the ‘average’ Pretest score.