Enhancing Model Visualization in Statistical Analysis: Introducing the R Package MoPlot

Umberto Granziol[1], Marianna Musa[1], Lorenzo Atzeni[2], Stefano Dalla Bona[1]

[1]Department of General Psychology, University of Padova, Padova, Italy; [2]Department of Developmental Psychology and Socialization, University of Padova, Italy

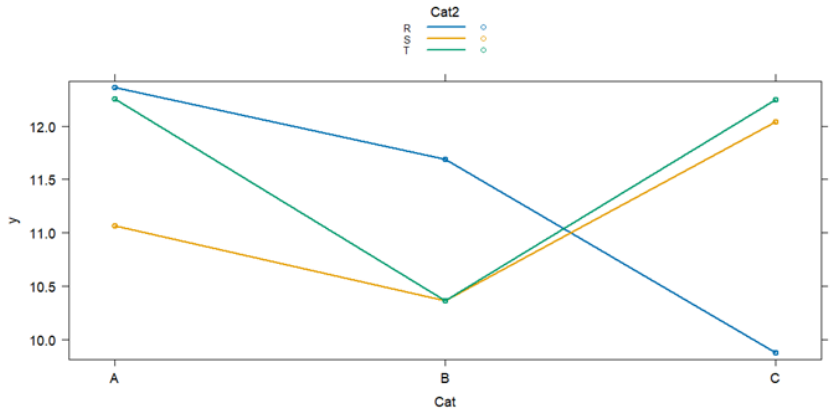
EAM 2025

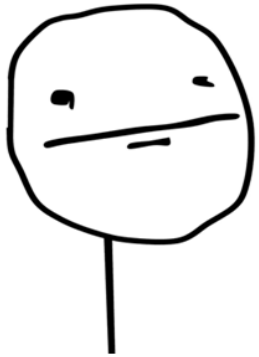
The complex art of model plotting

- Plotting model is an art (All the researchers I know, 2025).
- As any form of art, it takes time (and patience)..
- .. and sometimes can be difficult to understand.
- However, unlike art, graphs and their interpretation should not be left to the viewer.

The complex art of model plotting

It is so impressive how the artist shows the dynamism bringing from the A-B difference to B-C one, and how it is influenced by the second variable!





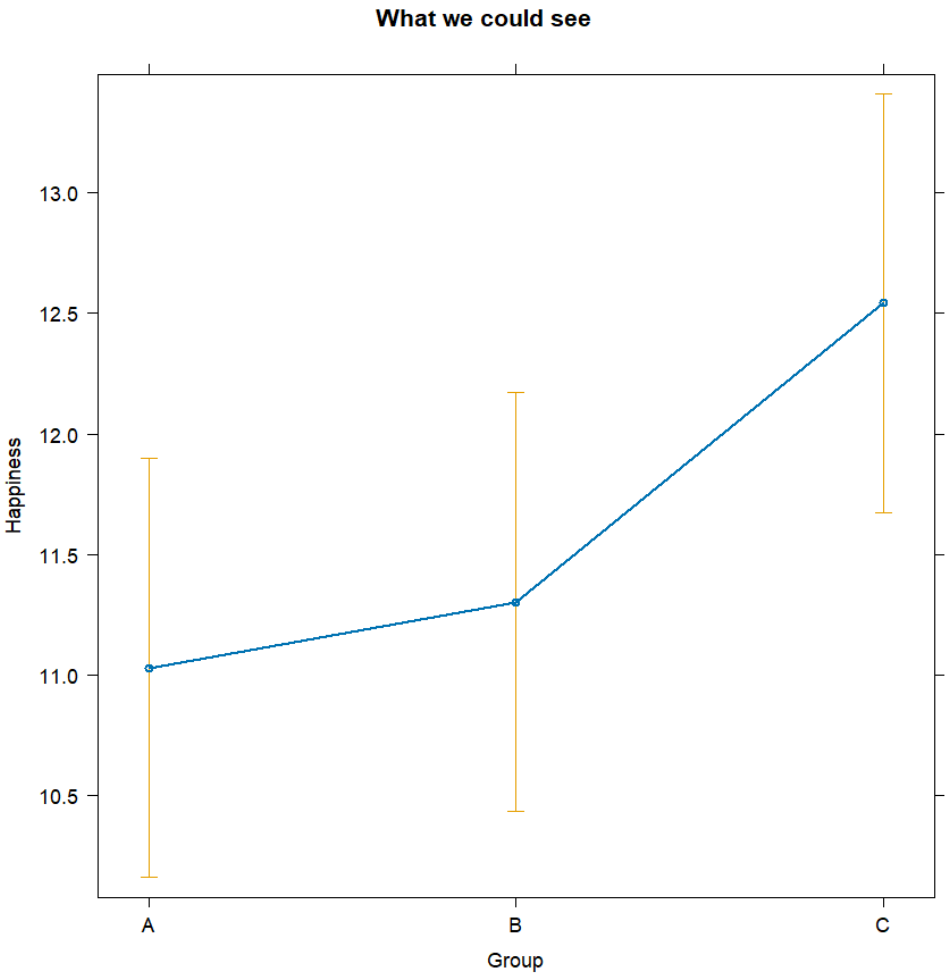
I do not know Phil..
I just see lines and dots...

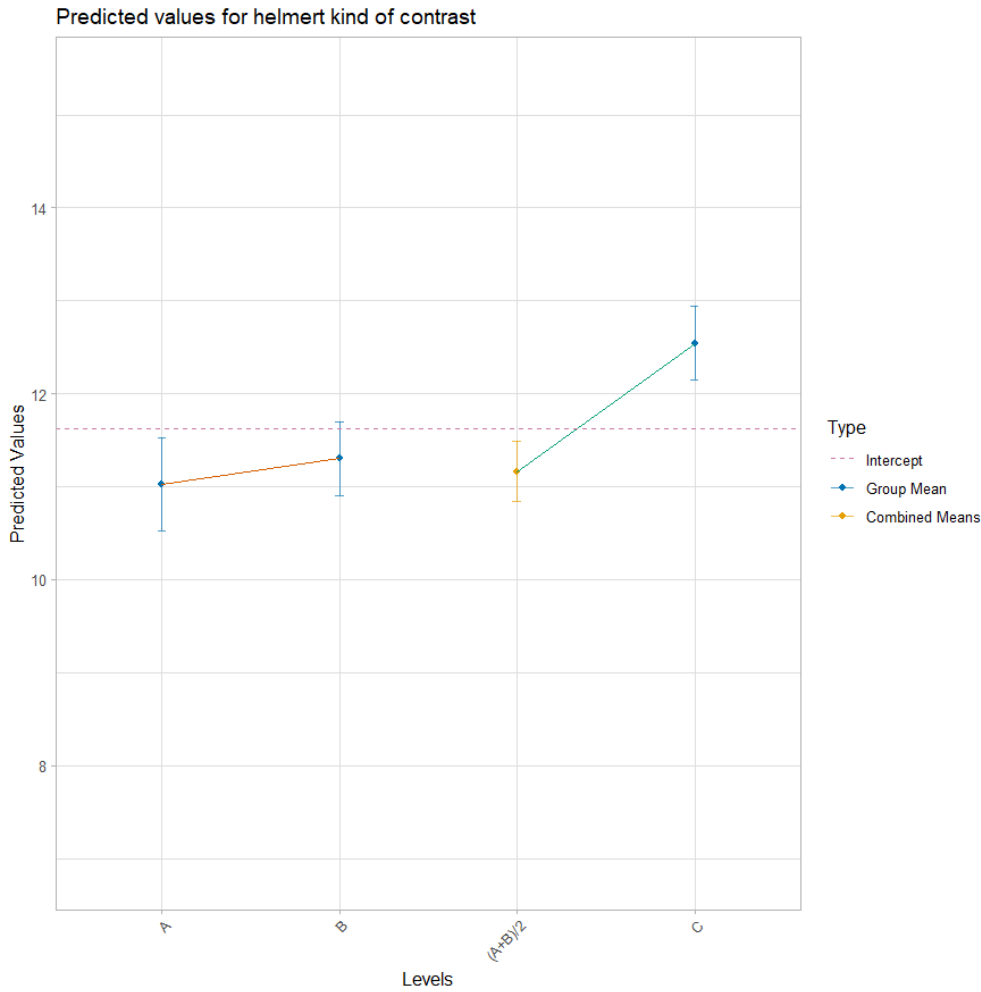
The complex art of model plotting

- Research question: Do researchers travelling for conferences feel happier than researchers who do not travel at all?
- Simple regression, categorical predictor (Group)
  1. A: No trip at all
  2. B: Travelling, but only for extra-uni business
  3. C: Travelling, a lot (for conferences?)
- Plannes comparisons: Helmert Contrasts (A vs B; A+B/2 vs C)

```
> summary(model)

Call:
lm(formula = Happiness ~ Group, data = prova)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1513 -1.0345 -0.0939  1.2057  4.2413

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.6245     0.2505  46.398   <2e-16 ***
Group1        0.1373     0.3068   0.447   0.6562
Group2        0.4592     0.1772   2.592   0.0121 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.941 on 57 degrees of freedom
Multiple R-squared:  0.1083,    Adjusted R-squared:  0.07698
F-statistic:  3.46 on 2 and 57 DF,  p-value: 0.03817
```

Reality *vs* Expectations





Problems and solutions

- Sometimes, pre-defined plots can be misleading, independently of the complexity of the model or the users' summary-reading skills.
- Such plots tend to simply show predicted means of all the possible levels.
- No consideration on planned comparisons/hypotheses.

So..

We introduce **MoPlot**, an R package that plots the comparisons and the relationships among variable found in a model, even in case of interactions (and further more).

How does it work?

INPUT & FUNCTIONING

- A statistical model.
- Specific variables (interactions)
- Type of variable.
- MoPlot examines the design matrix of the input.
- It checks the kind of contrast coding (for all variable).
- Extracts means/trends (raw or predicted) related to levels of all/specific model coefficients.

OUTPUT

- It provides two plots:
  1. A plot with means/or trends of target model coefficients (or all).
  2. A plot with model coefficients, with standardized parameters and effect size (Cohen's d, from `effectsize` package; Ben-Shachar et al., 2020)).
- It Respects the contrast coding used and the specific combinations defined by R model (one coefficient at time).
- Provides a ready-to-go caption for the plot, containing essential information on how to describe the plot.

Math & Stat Behind

MoPlot works with Design, Contrast and Hypothesis matrices:

- Assuming a linear model $Y = X\hat{\beta} + \hat{\epsilon}$:
  - Y is the response variable's vector
  - $\hat{\beta}$ is the set of expected regression coefficients
  - $\hat{\epsilon}$ is the expected error.
  - X is the design matrix of the model.
- In a design matrix, each row corresponds to a subject. The first column encodes the intercept. The other columns refer to each predictor.
- In case the predictor is a categorical variable, each column cell reflects the condition to which each subject is assigned and is usually coded with a number.
- In this way, a categorical variable can be handled as a numeric one by the linear model.

- Considering an example with nine subjects, one dependent variable y and one categorical predictor, coded with sliding difference contrasts:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \end{bmatrix} = \begin{bmatrix} 1 & -0.66 & -0.33 \\ 1 & -0.66 & -0.33 \\ 1 & -0.66 & -0.33 \\ 1 & 0.33 & -0.33 \\ 1 & 0.33 & -0.33 \\ 1 & 0.33 & -0.33 \\ 1 & 0.33 & 0.66 \\ 1 & 0.33 & 0.66 \\ 1 & 0.33 & 0.66 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \\ \epsilon_7 \\ \epsilon_8 \\ \epsilon_9 \end{bmatrix} \quad (1)$$

- The weights assigned to each row and coding the levels of a categorical variables can be summarized in the *contrast matrix* C.

The Contrast Matrix

- A contrast matrix C contains the groups or levels of a variable in the rows, and the specific comparisons to be tested are in the columns, along with their corresponding weights.
- C can be conceived as a way to adapt theoretical hypotheses into a form that is usable in the context of the linear model.
- The first column encodes the first contrast (or comparison).
- The second column encodes the second comparison.

$$C = \begin{bmatrix} -0.66 & -0.33 \\ 0.33 & -0.33 \\ 0.33 & 0.66 \end{bmatrix}$$

The hypothesis matrix

- Actually, the contrast matrix C is obtained from the hypothesis matrix H, by applying a set of matrix operations called generalized matrix inverse function ($C = (H'H)^{-1}H'$).
- In fact, the hypothesis matrix is a matrix H contains the hypothesis in rows and the levels of a variable in columns.
- Hypothesis matrices serve as valuable tools for conveniently representing particular hypotheses related to potential comparisons.
- In the case at hand:.

$$H = \begin{bmatrix} A & B & C \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

- All these elements can be found in our models. In R, for instance, the design matrix can be found using the `model.matrix()` function.
- Each column of this matrix will be use to obtain the correpsonding model coefficient.

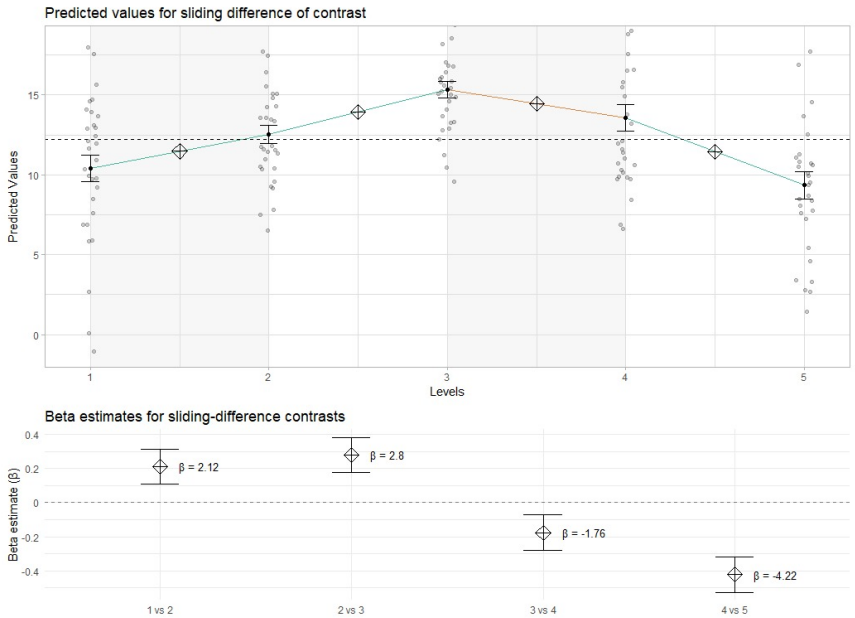
```
> model.matrix(model)
  (Intercept)     Cat2-1     Cat3-2
1           1 -0.6666667 -0.3333333
2           1 -0.6666667 -0.3333333
3           1 -0.6666667 -0.3333333
4           1  0.3333333 -0.3333333
5           1  0.3333333 -0.3333333
6           1  0.3333333 -0.3333333
7           1  0.3333333  0.6666667
8           1  0.3333333  0.6666667
9           1  0.3333333  0.6666667
attr(,"assign")
[1] 0 1 1
attr(,"contrasts")
attr(,"contrasts")$Cat
        2-1        3-2
A -0.6666667 -0.3333333
B  0.3333333 -0.3333333
C  0.3333333  0.6666667
```

```
> summary(model)

Call:
lm(formula = y ~ Cat, data = prova)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6627 -1.4334 -0.5391  1.6664  2.9887

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.3085     0.6829  18.024 1.88e-06 ***
Cat2-1       -3.9254     1.6728  -2.347   0.0573 .
Cat3-2        2.8117     1.6728   1.681   0.1438
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.049 on 6 degrees of freedom
Multiple R-squared:  0.4937,    Adjusted R-squared:  0.3249
F-statistic: 2.925 on 2 and 6 DF,  p-value: 0.1298
```

- Let's see how MoPlot uses such information.

Example 1: Single variable

- Simple regression, categorical predictor (vi, five levels).
- Sliding difference contrast coding.



Example 1: Single variable

```
> summary(modellino)

Call:
lm(formula = vd ~ vi, data = mydata)

Residuals:
     Min       1Q   Median       3Q      Max
-11.4858  -2.5132  -0.0736   2.4142  12.0860

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.2381     0.3277  37.348  < 2e-16 ***
vi2-1         2.1188     1.0362   2.045  0.04269 *
vi3-2         2.7997     1.0362   2.702  0.00772 **
vi4-3        -1.7642     1.0362  -1.703  0.09080 .
vi5-4        -4.2168     1.0362  -4.069 7.71e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.013 on 145 degrees of freedom
Multiple R-squared:  0.2288,    Adjusted R-squared:  0.2075
F-statistic: 10.75 on 4 and 145 DF,  p-value: 1.164e-07
```

Information/Caption

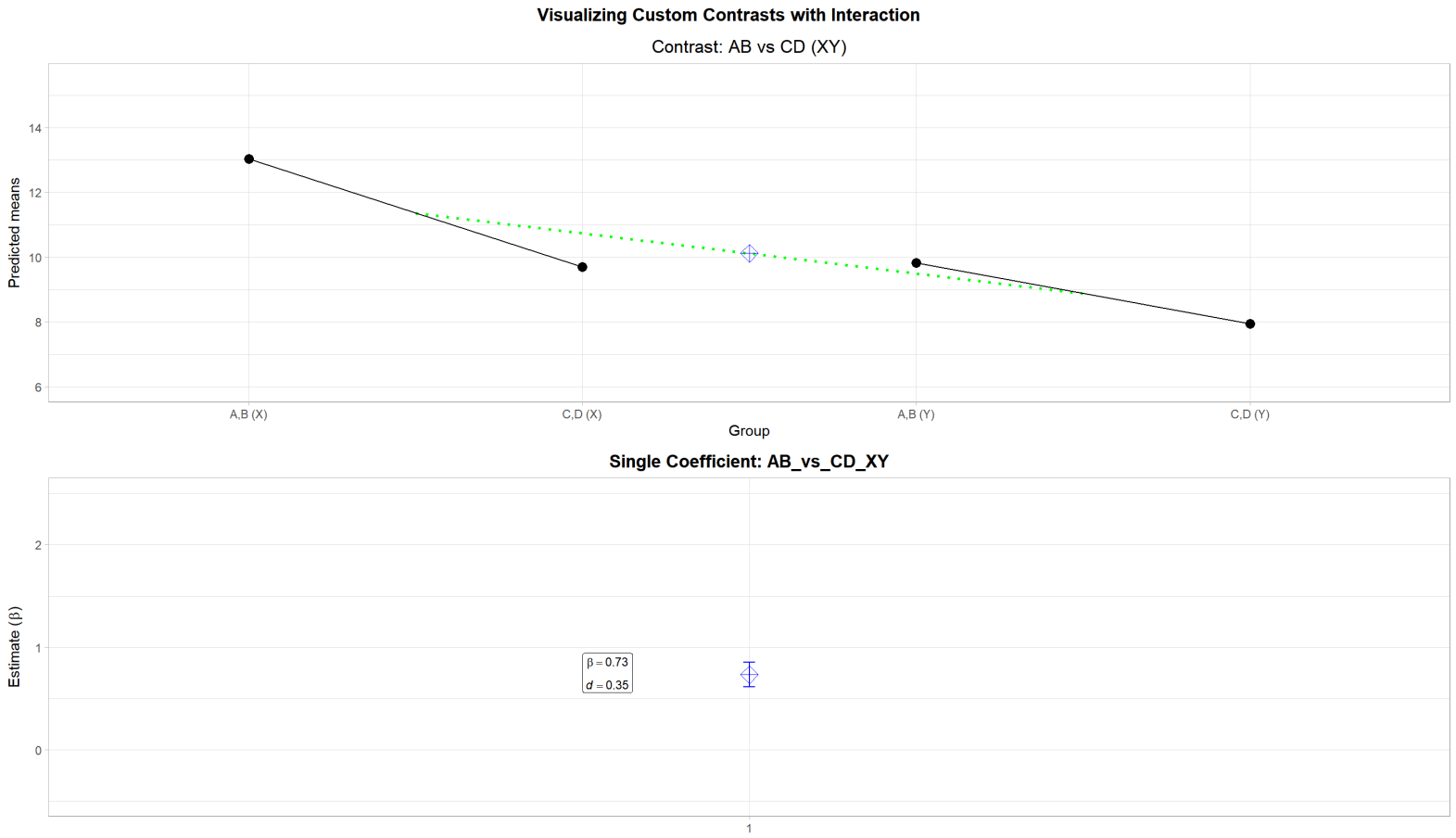
This graphical representation depicts the linear model, with vi as the categorical predictor (levels are represented on the x-axis), and vd as the numerical dependent variable (values are shown on the y-axis). The applied contrast type is sliding difference where each group mean is compared to the mean of the subsequent group. Blue dots represent the expected values (means) for each group, and error bars indicate the uncertainty associated with these estimates. Error bars indicate the uncertainty associated with each expected value. The dashed purple line marks the baseline expected value. Green lines highlight significant contrasts, with 0.05 serving as the threshold for the first type of error.

Example 2: interactions

- Multiple regression, two categorical predictors.
  1. Label1 (four levels: A, B, C, D). Customized contrasts (i.e., A+B/2 *vs* C+D/2).
  2. Label2 (two levels: X, Y). Sum contrast coding.

```
> summary(model)

Call:
lm(formula = y ~ Label1 * Label2, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9673 -1.4484  0.0064  1.2811  5.2007

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      10.1227     0.1487  68.079  < 2e-16 ***
Label11           2.6132     0.2974   8.788 7.71e-16 ***
Label21           1.2400     0.1487   8.340 1.30e-14 ***
Label11:Label21   0.7322     0.2974   2.462   0.0147 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘

Residual standard error: 2.096 on 196 degrees of freedom
Multiple R-squared:  0.4491,    Adjusted R-squared:  0.4407
F-statistic: 53.27 on 3 and 196 DF,  p-value: < 2.2e-16
```

Example 2: interactions



- The lines (upper plot) connecting the mean differences (over the predicted means) help to understand how to obtain the regression coefficients (lower plot).

Limitations == Future directions

- This is a preliminary version of the package.
- So far, MoPlot supports linear models.
- We are currently implementing the interactions with continuous variables.
- Coming soon: implementation for glm,lmm, glmer.
- Comments? Ideas? Highly appreciated!

Trust is good, contrasts are better

"So it's time to leave you a preview,
so you too can review what we do.."
(Harder than you think, Public Enemy; 2007)

Thank you for youR attention!