We select randomly 1/2 of the total number of participants and we divide them in 2 sub-groups.  We provide two trainings of 40-120 mins, one group attends GRL training, one ArgGRL.

We give 40 mins training. Afterwards we let each group ask questions and continue the training as long as it is needed ==> *we measure the necessary time to get acquainted with each framework; we capture the questions, therefore the difficulties to learn/use one over the other*

We select two persons from each sub-group, they become the **modellers** of the experiment, the rest become the **analysts**.

We conduct 2 parallel sessions, each with two **modellers** plus 1/2 of the **un-trained participants**.

We give each group of **un-trained participants** the sametask (for example, define goals and requirements for an e-learning platform). The **modellers** have the individual task to attend and capture their session and transform it in a GRL/ArgGRL model. Modeller can take notes and record the session and then create the models.

Time could be 15 mins for discussions and 45-60 mins for the creation of the models.

==> *we masure the details/information captured in the models, ArgGRL vs GRL*

We use two modellers so we can pick the more detailed model/ we can combine and provide further a more refined version.

We give each group of  **analysts,** one the GRL model and the second the ArgGRL model. We ask them to identify the goals and reasoning from the models. They also have to write a list of ambiguities, and problematic points.

==> *we measure the accuracy of the interpretation vs. the intended design;*

*==> we measure how many elements were identified correctly using each framework*

*==> we measure how many ambiguities and problems were identified by each group*

**Comments Floris**

As for the experiment itself, my main worry is how we will measure the following

==> *we measure the accuracy of the interpretation vs. the intended design;*

*==> we measure how many elements were identified correctly using each framework*

*==> we measure how many ambiguities and problems were identified by each group*

Can you explain precisely how this works? What is "correctly identified"? What is "accuracy"? And: can you come up with an example problem that is interesting enough (i,e. no toy example), but small enough to be done in a few hours? Note that the Sing et al paper about argumentation and goals is based on a pre-prepared dataset, and the measurements are also based on that.