We thank the reviewers for their careful consideration of our paper and for the useful feedback. We are happy to see that the reviewers find that the paper is “*well-written*”, contributes a “*cool, instructive result*”, and that it “*tackles a very important question on representation learning and provides interesting new insights about it*”. However, there appear two separate major concerns by R1 and R2 to which we kindly respond below.

About the Abstract Visual Reasoning Task (R1)

R1 raised the concern that the abstract visual reasoning tasks considered in this paper “. . . *seem a bit unintuitive and overly difficult*”, finds it “. . . *weird the authors didn’t just consider a task with one row and one panel missing and the same one factor changing between panels*”, and requires “. . . *a very good explanation as to why the strange formulation in the paper was used and simpler tasks weren’t used*”. We will provide this explanation next.

- **Abstract visual reasoning to evaluate disentangled representations.** Disentanglement prescribes *how* information about salient factors of variation in the input should be encoded. Hence, we require a task that (1) involves reasoning about these factors, (2) that is well established and connects to the literature, and (3) that can not be trivially solved through correlating image statistics. This leads us to Raven’s Progressive Matrices (RPMs; [59]).
- **RPMs are a standard test for human abstract reasoning.** 3×3 RPMs (standard setting - as in our paper) require one to identify relationships among salient factors of variation in two complete rows of images, and apply these relationships to answer a multiple choice question to complete the final panel of the third row. RPMs are a standard test to estimate abstract reasoning capabilities (Motta & Joseph, *Handbook of Psychological Assessment*, 2016).
- **RPMs are used in prior work in ML.** RPMs have been considered in prior work in ML, e.g. [30, 63, 68], and in the cognitive science literature, see (Lovett & Forbus, *Modeling visual problem solving as analogical reasoning*, 2017) for an overview. We created RPMs on two disentanglement data sets as closely as possible to the prior work.
- **RPMs cannot be solved trivially.** In recent work [63] the abstract reasoning capacity of several deep neural networks were tested on a data set consisting of 3×3 RPM-like abstract visual reasoning tasks. It was confirmed that this is a difficult task at which traditional architectures struggle, while the WReN architecture (that incorporates elements to enhance relational reasoning capabilities) performs best.
- **Validated through initial study & humans.** Our initial study in Section 4.2.1 validates that our adaptation of the RPM task serves as a sensible benchmark for disentangled representations, and we have further informally tested the task with >10 people. Finally, we note that we only consider the AND relationship, which is a simpler setting, and that the data sets considered in fact test two difficulties due to differences in their number of possible feature combinations.

About Methodological Concerns (R2)

R2 raised concerns about the methodology in this paper, arguing that “*If ground-truth factors are available, then we can directly use the ground-truth factors to train WReN and achieve excellent performance . . . but we do not need disentanglement learning*, and that “*if ground-truth factors are not available, then we can not compute disentanglement scores, and we are not able to utilize the results are shown in Figure 3, 4 and 5 to select the best disentangled representation”*. It appears that there is some confusion with regards to the goals and contributions of this paper.

- **Relevance: validate the motivation of >20 recent ML papers.** Recently, numerous papers have been concerned with learning disentangled representations [1, 8, 9, 10, 11, 15, 16, 17, 18, 20, 26, 27, 28, 29, 35, 42, 47, 51, 55, 56, 60, 61, 62, 76, 77]. The key motivation (but also assumption) of these works is that current notions of disentanglement (MIG, DCI, *etc.*) are desirable but until now there has been little empirical evidence verifying this.
- **How? By evaluating disentangled representations.** We are hence concerned with *evaluating the usefulness* of disentangled representations (for abstract visual reasoning), rather than *learning* disentangled representations. This distinction is critical: We do not make any assumptions about the *feasibility* of (and *methods* for) learning disentangled representations in the absence of ground-truth factors, and our results have implications for the supervised, semi-supervised, and unsupervised settings. In particular, our results highlight the benefits that current and future research on disentanglement may provide for solving non-trivial upstream tasks that require abstract reasoning.
- **Originality and contribution: novel experimental setup with non-trivial insight.** The research question and the experimental setup is novel and lead to novel insights. Notably on two relevant and non-trivial abstract visual reasoning tasks we find that disentangled representations enable quicker learning using fewer samples. Compare this to a most recent critical work [50], where it could not be observed that higher disentanglement scores reliably lead to a higher sample efficiency on a simple upstream single-factor classification task.

Additional comments

We will release all data sets, pre-trained models and code upon publication. Following R3s suggestions we will further improve upon the presentations of results, captions and visualizations.