Overall, my evaluation of this proposal in terms of the five Stage 1 criteria is favorable. In my assessment the proposed research is assessing a scientifically valid question, with an adequate rationale behind the hypotheses being tested, and a basically sound study design and analytic plan that has been explained with sufficient detail to allow other investigators to conduct close-to-exact replications of this work and prevent undisclosed flexibility in the analysis of the resulting data. For Criteria B – E, I did perceive areas with room for improvement, or at least where the current state of the proposal raised questions that I would encourage the authors to consider as they continue to plan this work. I have outlined these thoughts under the heading of each criterion below. All of these issues are quite minor in scope, and do not seriously compromise the basic soundness of this proposal.

In the interest of transparency, I would like to acknowledge that this is the first registered report I have reviewed – so forgive me if my comments are disorganized, insufficiently clear/precise, or not germane to the RR context. I am happy to communicate further with the authors if that would be helpful!

Reviewer: Andrew G. Christy, Ph.D. (Assistant Professor, SUNY Plattsburgh; achri021@plattsburgh.edu)
Signature:

* **Criterion 1A: The scientific validity of the research question(s).**
	+ The proposed study will test effects observed in prior research; thus there is an existing evidence base on which the question and hypotheses rest.
* **Criterion 1B: The logic, rationale, and plausibility of the proposed hypotheses.**
	+ **Rationale for exploratory tests of social norms could be explained more clearly on pg. 13**
	+ **Rationale given on pg. 14 for adding “independent” and “other” options to political-orientation measure could also be clarified – it seems that the words *do not* were omitted (these additions can represent people who *do not* identify as Republicans or Democrats).**
		- **“conservatives or liberals” would be a clearer phrase to use in that sentence also – truer to the content of the item, and clearer as a label for political ideologies.**
	+ **Rationale for including the general lay beliefs about good true selves could be clearer (currently no rationale is given, it is simply stated that this measure will be included). If it is being included for purely exploratory purposes (which is what it seems like), that could be made explicit. If the authors intend to test hypotheses related to that variable (even tentative hypotheses), those hypotheses and an accompanying rationale should be stated.**
* **Criterion 1C: The soundness and feasibility of the methodology and analysis pipeline.**
	+ **The phrasing of the “surface self” response option implies that everything a person learns (i.e., qualities of the person resulting from “nurture” influences) are by definition not part of that person’s true self. That doesn’t seem like the most neutral way to measure that concept; while on average a person’s innate qualities might be perceived as more authentic/essential than their acquired qualities, that is not entailed by the bare concepts of the true vs. surface selves. The phrasing of the “true self” response option (“the deepest, most essential aspect of this person’s being”) is more neutral – that phrasing doesn’t directly imply innateness. Can the “surface self” response option be revised to more closely parallel that language? For example, “This person’s ‘surface self’ (the shallower, non-essential aspects of this person’s being).”**
		- **There might be some risk that this phrasing (essential vs. non-essential) is not readily understood by all participants, which could introduce noise/imprecision into the measurements. With that in mind, it may be worth considering more accessible alternative phrasing that preserves the same basic meaning. For example: “The person’s ‘true self’ (who the person really is, deep down)” vs. “The person’s ‘surface self’ (parts of the person that don’t define who they really are).” I’m not convinced this phrasing is perfect, just an example of how these items’ wording might be made more accessible.**
		- **I recognize that these suggestions entail further deviations from the original work being replicated here. In making these suggestions, I am prioritizing considerations of precise and valid measurement of the relevant constructs. Those considerations seem more important than fidelity to the original studies – but I respect the authors’ right to disagree with that assessment and opt to stick as closely as possible to the original study materials!**
	+ **On a similar note – the wording of the true-self rating items in the Study 2 replication is also rather convoluted and potentially confusing. In particular, the word “oneself” in those items is strange – just saying “them” would be clearer (e.g., “…there was always something deep within *them* calling this person…”). But the overall phrasing and structure of these items is odd and seems likely to confuse at least some participants. Alternative phrasing like: “Now that they have [CHANGE], this person is now…” with a response scale indicating degrees of closeness/farness from being their true self (e.g., 1 = *much farther from being their true self*, 4 = *no closer to or farther from being their true self*, 7 = *much closer to being their true self*)**
		- **The same considerations apply as above – if the authors judge that the wording should remain relatively unchanged in the interest of fidelity to the original work, that is reasonable. But in that case I would still advocate for changing “oneself” in these items to “them”!**
	+ **The reasoning behind the power analyses (which the authors specifically requested feedback on) seems adequate to me. Even though not all of the original effects could be estimated precisely, the planned sample size is adequate to detect much smaller effects than those originally reported. In my own experience doing similar research, these true-self/morality effects are pretty robust – so I believe statistical power will be more than adequate.**
	+ **In the consenting process in the Qualtrics survey, it states that the study will include attention/comprehension checks. However, it doesn’t seem that any checks are actually included. While there are quite a few safeguards in place to filter out bots and non-serious participants, it does seem like including some simple instructional attention checks embedded throughout the study, and/or simple comprehension checks (e.g., a single item accompanying each vignette asking participants to identify the nature of the change the target person underwent) could add value when it comes to ensuring data quality. The latter type of items (comprehension checks) would probably be sufficient, since they effectively double as attention checks. Including these checks would likely result in a more exclusions. This could be offset by setting a relatively high bar for exclusion (i.e., participants would only be excluded if they displayed a persistent pattern of inattentiveness/non-comprehension across multiple check items). Increasing the planned sample size further could also offset a larger number of anticipated exclusions, though I’m not sure that’s necessary. The pre-registration plan could also be updated to apply these exclusion criteria separately from the others (i.e., the plan would include three stages of analysis: 1. pre-exclusion data, 2. data after applying the currently proposed set of exclusion criteria, 3. data after applying the full set of exclusion criteria including these new check items).**
	+ In the supplemental materials, it is stated that recruitment on MTurk will be limited to workers who have completed between 5,000 and 100,000 HITs. That strikes me as odd; only highly experienced MTurk users would be eligible for participation in this study, and more casual users would be excluded. Allowing less experienced MTurk users to participate would seem desirable for a number of reasons (i.e., greater proportion of naïve participants, somewhat more representative of people in general).
* **Criterion 1D: Whether the clarity and degree of methodological detail is sufficient to closely replicate the proposed study procedures and analysis pipeline and to prevent undisclosed flexibility in the procedures and analyses.**
	+ **The exclusion criteria described in the manuscript do not match exactly with those outlined in the supplemental material (commenting on this here since it seems to be related to potential flexibility in analyses – but some of my comments under the previous criterion are also related to this).**
* **Criterion 1E: Whether the authors have considered sufficient outcome-neutral conditions.**
	+ **Manipulation checks were added to the Study 1 replication, but not in the Study 2 replication. It would seem possible to add a similar morality rating there, allowing the authors to test whether the “pro-liberal” and “pro-conservative” changes were indeed perceived as changes for the better by liberal and conservative participants, respectively.**