The OPEN Bilingual Advantage Replicability project (OPEN BAR) – Original project proposal

Giacomo Spinelli

State of the art and objectives

State of the art

No one doubts that **bilingualism (broadly defined as speaking more than one language)** comes with cultural, social, and economic advantages. Yet, seldom has an idea spurred a more heated debate than the idea that those advantages would extend to cognition—specifically, to **attentional control (AC)**, a set of abilities involved in controlling not only verbal but all types of information [1]. Known as the **Bilingual Advantage (BA) hypothesis**, the idea has gone, in the span of two decades, from universal acclaim to increasingly skeptical scrutiny [2].

Intense scrutiny was inevitable given the **tangible relevance a cognitive advantage would have in one's life**, especially when the relevant abilities are being developed or risk deterioration (e.g., with aging). For this reason, the debate has been largely restricted to alleged BAs in measures for which such an advantage can be directly adjudicated—**overt behavior and neural efficiency in performing an AC task** (e.g., Stroop)—with far less contention for monolingual-bilingual differences in measures not indicative of a potential advantage per se (i.e., without the advantage being corroborated by a measure of the former type), such as neuroanatomical/plasticity ones (measures that, being important but marginal to the discussion, will not be considered further) [3]. But why the skepticism?

Reasons for skepticism on the BA hypothesis

The primary reason is empirical: Early BA reports have gradually given way to a wave of null results, with the average effect size in recent meta-analyses being small at best and **indistinguishable from zero when corrected for publication bias** [4]. Worse, some meta-analyses have shown larger effect sizes with studies with smaller samples and/or poorer quality, suggesting a potential involvement of **questionable research practices** [5].

But the reasons for skepticism are also theoretical. There is recurring **vagueness in the assumptions and predictions** made by the various versions of the BA hypothesis [6]. Indeed, although current BA theories recognize more nuance in bilingualism and its effects on cognition, they have also substantially extended the range of processes in which such effects could emerge—including, e.g., "facilitation of relevant processing", a process involved in virtually any experimental task [1].

Concerns with such theories thus seem to be well justified, and **I've played a leading role in voicing them**. For example, noting that some BAs are such only based on **ad hoc interpretations**, I've shown such interpretations do not hold when more theoretically grounded manipulations are used [7]. I've also pointed out the risk of **circular reasoning** associated with current BA theories [8] and demonstrated that the unverified assumptions those theories rest on (i.e., that bilingualism involves some form of AC) completely collapse when put to test [9–10].

In sum, the BA hypothesis appears to have a lot less going for it than it initially appeared, inducing some of its most vigorous opponents to describe it as a **textbook example of the broader**

replication crisis affecting natural and social sciences [3]. I contend, however, that this description is not entirely accurate.

A case of replication crisis?

The BA hypothesis certainly shares some similarities with notorious cases of replication crisis such as social priming—the once-popular idea that behavior can be unconsciously influenced by brief exposure to subtle cues [11]. Social priming had accrued most of its credibility thanks to successful **conceptual replications**—studies that followed up on previous studies by attempting to extend their findings to new measures or situations—before being brought down by failed **direct replications**—studies that reproduced previous studies as closely as possible with no attempt at extension [12].

Similarly, many of the positive results in the BA literature have come from studies that attempted to extend the seminal effects to new bilingualism and/or AC measures, with null results most typically coming from studies not aiming at extension [3]. The first class of studies naturally fall under the umbrella of conceptual replication, creating an obvious problem because **conceptual replication success interacts with publication bias** (i.e., a successful conceptual replication is more likely to be submitted and get published than an unsuccessful one). As a result, by their very own nature, conceptual replications tend to give prominence to false positives and ultimately create the **distorted image of a robust phenomenon**—what is thought to have happened with social priming [12].

What makes the BA case different from textbook cases of replication crisis, however, is that the second class of studies—those not aiming at extension, which most typically produced null results—**have often come short of meeting the standards of direct replication**. That is, while attempting to maintain general aspects of previous studies (e.g., the use of bilingualism and AC measures), many studies have also introduced some deviations (e.g., the specific materials used to measure bilingualism/AC or the preprocessing/analysis routines). Albeit seemingly small, those deviations have been sufficient for BA supporters to **dismiss most of the null effects produced by those studies**, reasoning that many variables likely modulate the emergence of BAs and null results are to be expected when changing them [1]. With skeptics viewing such arguments as a mere demonstration of BA's fragility, the end result is a stalemate in which, to this day, the two factions keep pursuing their research without reaching **any sort of consensus**—a situation much different from that of other replication crisis cases, which appear to have taken a turn towards resolution, e.g., [13]. How can the BA debate also take such a turn?

Towards a resolution of the BA debate

The solution is apparently simple: **direct replication**. Directly replicating previous studies reporting a BA by taking care to reproduce them as closely as possible addresses any potential counterargument in case of failure and is useful either way [14]: If the key outcome is successfully replicated, it increases confidence that the original result was a true positive, as BA supporters maintain; if it isn't, it increases confidence that it was a false positive, as skeptics suggest. Direct replication thus provides **a more informative and objective way to establish the replicability of a phenomenon** than do non-direct replications.

If the solution was that simple, however, why have we not used it already? In addition to general challenges with direct replication (e.g., with securing the information and resources necessary for independent reproduction of another lab's study, or with the incentives to conduct a study likely to produce a null result [12]), **BA studies might be inherently harder to directly replicate** than others because they typically include multiple measures for which potentially meaningful deviations from the original study can arise, including those relative to the population sampled (which are hard to reproduce without direct access to the same population).

These challenges, however, are not insurmountable given the latest advances in scientific practices. For example, with **crowdsourcing** having become a **widespread and reliable means of data collection** [15], it is possible to collect samples with specific characteristics faster and in greater quantity than in traditional lab studies. Further, the practice of sharing the information and resources necessary to reproduce a study in **public repositories** is on the rise, although many still do not engage in such practices or do so incompletely [16]. In addition to the possibility for researchers to individually preregister a study, more and more journals also offer the option of **registered reports** as a way to evaluate the scientific validity of a study independently from its results, contrasting existing incentives towards publication bias and questionable research practices [17].

Such advances have created **fertile soil for direct replication to thrive**. Indeed, direct replication has been the unifying factor for several **multi-lab initiatives** that have greatly contributed to our understanding of the strengths and weaknesses of contemporary science [18–19] (a type of initiative **I've been an active part of in past [20] and ongoing work**). These initiatives and related ones, such as registries for replicability tracking [21] and community-augmented meta-analyses (CAMAs—a continuously updated, interactive type of meta-analysis based on community input [22]), are indicative of a shift towards **a more dynamic, collaborative way to do science**.

In the BA literature, calls for this type of shift have repeatedly been made [6, 23] and fallen unheeded**.** The problem is the lack of support for such a shift**.** Specifically, **I see a gap** between individual-study repositories (which are sparse and lack standardization) and generalist replication registries (which lack focus on specific domains) that only a **BA-specific tool designed for direct replication** would be able to fill. With such a tool, the BA community would be able to determine which effects are real and which are not and move the debate towards a constructive resolution. Still, albeit necessary, such a tool would not suffice to address more specific challenges that might contribute to BAs' low replicability.

Make the unreplicable replicable

Among the challenges for replicability of BA studies, **insufficient difficulty of AC tasks** and **exclusive focus on behavior** have been argued to be leading causes of failure to observe BAs [1]. These arguments are not without problems (e.g., BAs have been observed in many relatively easy conditions [3] and inferring a BA from neural measures when there is none in behavior has questionable validity [24]). However, they do suggest that, for a BA to be highly replicable, AC tasks should: 1) be demonstrably difficult for participants, ideally without altering the basic processes they involve (e.g., an AC task with many—as opposed to few—stimuli and responses [25]); 2) involve both behavioral and neural measures (e.g., **event-related potentials (ERP) from the electro-encephalography (EEG) signal** [24]).

Perhaps more pressing still are the challenges with the **psychometric properties** of the measures involved in BA studies. It is now well known that AC measures suffer from **poor reliability** (consistency in measurement) [26]. Treating such a problem without altering AC measures (and their interpretability) requires reducing trial variability by both dramatically increasing trial size and accounting for such variability in the analyses [27], practices that remain uncommon. Similar problems are starting to arise for language background questionnaires—the typical bilingualism measures used in BA studies. Specifically, those measures are subjective [28] and pose retrospective questions hard to respond to without imprecision or bias [29], problems that do not apply to objective measures (e.g., standardized proficiency tests) and real-time sampling of individuals' experience (e.g., ecological momentary assessment (EMA) [30]). As they stand, however, typical AC and bilingualism measures are not posed well to succeed because low reliability of one or, worse, two variables **inevitably attenuates the observable correlations between them** [26].

A less appreciated psychometric problem is with the **construct validity** of AC tasks (their accuracy in measuring AC itself): AC tasks may not correlate well not only because of their low reliability but also because of their improper use. For example, vocal- and manual-response versions of the color-word Stroop task are often used interchangeably even though they belong to different task ensembles (Stroop and Stroop-like, respectively), with a similar situation arising for spatial Stroop and Simon tasks (Stroop and Simon task ensembles, respectively) [31]. **My extensive work at the forefront of research with AC tasks** (e.g., [32–34]) has shown that these differences can and do play a role in terms of the abilities involved, likely limiting the degree to which the tasks' measures map onto the same construct and, hence, the observable correlations between them. From this work also follows, however, that construct validity should improve with a theoretically clearer and carefully matched set of tasks. In sum, replicability of BA studies faces a number of challenges, and, while some BA studies have certainly addressed at least one of them, **none has addressed all of them at once**.

Objectives and overview of activities

OPEN BAR has three objectives, each mapped onto a separate milestone and work package. The first is **enabling replicability** (WP1), i.e., empowering researchers to perform direct replication of BA studies with a dedicated tool. To that aim, I will develop a **community-augmented web app** with a large initial set of successful BA studies complete with the information and resources necessary for direct replication presented in an easy to use, standardized format. Users will be able to interact with the app and contribute to its extension, with a reward system in place to motivate engagement. The app will be then further extended with tools for sample planning, automatized study addition, and dynamic meta-analysis, making it an **all-rounded tool for BA research**.

The second objective is **estimating replicability** (WP2), i.e., quantifying the direct replicability of BA studies with the **first large-scale, multiple-study direct replication project** in the literature. A total of 18 studies (including two EEG ones) will be selected for direct replication in a **registered report format** involving feedback from the original authors themselves. Reproducing those studies as closely as possible but with larger sample sizes to account for greater power to detect the original effects, I will leverage cutting-edge meta-analytic practices to examine several indicators and moderators of replication success (including preprocessing/analytical choices). No matter the

outcome, this work will demonstrate the usefulness of direct replication and provide the **springboard for more global replication endeavors**.

Finally, the third objective, **enhancing replicability** (WP3), refers to developing a type of BA study that addresses at once all known threats to BA replicability by: 1) examining both behavioral and neural (EEG) measures of AC performance; 2) creating an effective, yet not process-altering difficulty in AC tasks; 3) using theoretically clear and well-matched AC tasks; 4) increasing precision in both AC tasks (with a large trial size and treatment of nuisance effects) and language measures (supplementing standard questionnaires with both objective proficiency tests and EMA). Two studies will be conducted, the first with a piloting function and the second with an **unprecedented combination of trial and sample size** allowing a thorough examination of the robustness of any emerging BA. Regardless of whether a BA will in fact emerge, the study will be **a milestone in the BA literature for the quality and quantity of data provided**.

References

1. Bialystok & Craik (2022). https://doi.org/10.3758/s13423-022-02057-5
2. Antoniou (2019). https://doi.org/10.1146/annurev-linguistics-011718-011820
3. Paap (2022). https://doi.org/10.4324/9781003308027
4. Paap et al. (2024). https://doi.org/10.1177/13670069231214599
5. John et al. (2012). https://doi.org/10.1177/0956797611430953
6. De Bruin et al. (2021). https://doi.org/10.1162/nol_a_00038
7. Spinelli et al. (2022). https://doi.org/10.1037/xlm0001095
8. Spinelli & Sulpizio (2025). https://doi.org/10.31234/osf.io/2ah5t_v1
9. Spinelli & Sulpizio (2024). https://doi.org/10.3758/s13423-024-02503-6
10. Spinelli & Sulpizio (2025). https://doi.org/10.3758/s13423-025-02672-y
11. Molden (2014). https://doi.org/10.1521/soco.2014.32.supp.1
12. Harris et al. (2021). https://doi.org/10.1080/1047840X.2021.1889317
13. Cesario et al. (2017). https://doi.org/10.1080/23743603.2017.1309876
14. Nosek & Errington (2020). https://doi.org/10.1371/journal.pbio.3000691
15. Zwaan et al. (2018). https://doi.org/10.3758/s13423-017-1348-y
16. Ferguson et al. (2023). https://doi.org/10.1038/s41467-023-41111-1
17. Lakens et al. (2025). https://doi.org/10.1080/2833373X.2024.2376046
18. Camerer et al. (2018). https://doi.org/10.1038/s41562-018-0399-z
19. Open Science Collaboration (2015). https://doi.org/10.1126/science.aac4716
20. Bognar et al. (2024). https://doi.org/10.1038/s41598-024-53090-4
21. Röseler et al. (2024). https://doi.org/10.5334/jopd.101
22. Tsuji et al. (2014). https://doi.org/10.1177/1745691614552498
23. Leivada et al. (2020). https://doi.org/10.1017/S1366728920000358
24. Cespón & Carreiras (2020). https://doi.org/10.1016/j.neubiorev.2020.07.030
25. Donohue et al. (2016). https://doi.org/10.1016/j.neuropsychologia.2016.01.035
26. Hedge et al. (2018). https://doi.org/10.3758/s13428-017-0935-1
27. Lee et al. (2023). https://doi.org/10.31234/osf.io/rj2bu
28. Tomoschuk et al. (2019). https://doi.org/10.1017/S1366728918000421
29. Jylkkä et al. (2020). https://doi.org/10.1017/S1366728918001190
30. Shiffman et al. (2008). https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

31. Viviani et al. (2024). https://doi.org/10.3758/s13428-023-02215-0
32. Spinelli et al. (2019). https://doi.org/10.1037/xhp0000656
33. Spinelli et al. (2022). https://doi.org/10.3758/s13423-022-02138-5
34. Spinelli & Lupker (2023). https://doi.org/10.1037/xlm0001144

Research Methodology

OPEN BAR is articulated in three WPs corresponding to the project's objectives, with the completion of each constituting a milestone: enabling (WP1), estimating (WP2), and enhancing replicability (WP3). Resting on a synergy between open, meta-, and cognitive science, OPEN BAR is a **highly ambitious endeavor**, yet one I designed to remain **feasible**: With **minimal interdependence between the WPs**, I will be able to pursue them in parallel with little risk that potential hiccups in one would interfere with the others; and with an **excellent interdisciplinary team** of three postdocs and two PhD students, I will guarantee expert and adequate workforce to achieve all of the objectives, even where my knowledge is more limited. My profile as a **leading voice in the BA debate** and the affirmed position of Sapienza University's Psychology Department (**"Dipartimento di Eccellenza" 2023–2027**) further boost OPEN BAR's success potential.

WP1: Enabling replicability

In WP1 I will create the OPEN BAR app, a **community-augmented web app designed for direct replication of BA studies**. There will be three main activities. In WP1.1, I will gather information from existing BA studies for inclusion in the app. First, a list of eligible studies obtained from available meta-analyses and through a supplementary literature search will be annotated for study type (behavioral/EEG/fMRI/other), AC domain(s), age group, and outcome (bilingual advantage, disadvantage, null, or unclear). From this list, ~100 studies will be selected for complete information/resource extraction from their manuscripts, supplementary materials, and (if needed) original authors, prioritizing studies reporting a BA and representativeness in terms of type, domain, and age group. Aiming for **maximal comprehensiveness**, I will attempt to retrieve all useful aspects for reproducing the study, including those not typically shared (e.g., original study scripts).

In WP1.2 I will create the app itself. The app will come with an **easy to navigate graphical user interface** and feature increasing privileges based on user status. Unregistered users will be able to view/download all collected information/resources for all listed studies—including "to-do" ones—presented in a **standardized format abiding by the FAIR principles** (defined through a dedicated **data-management plan**) [1] and linked to their associated journal article, preprint, and/or data repositories.

Registered users, authenticated via their academic credentials (e.g., ORCID), will also have editing privileges for individual study entries based on their author/non-author status. For example, pending verification, non-authors will be able to add missing information/extra resources (e.g., alternative-software versions of original scripts) to existing studies or add new studies altogether (including ongoing ones and direct replication attempts). In addition, authors will be able to directly add/edit all information/resources to their own studies and provide public feedback on their accuracy and that of their direct replications. **Engagement will be motivated** by quantifying the completeness of individual study entries with a reproducibility index, with more reproducible studies being awarded badges of increasing symbolic value as proxies of research quality [2].

Following launch, in WP1.3 I will extend the app further with: 1) a power-analysis tool to plan the sample size needed to directly replicate a listed study based on its effect size or a fraction of it; 2) automatized addition (subject to community verification) of new BA studies as they are published based on their indexed information; 3) an interface capable of both tracking replication success [3] and perform meta-analyses interactively based on user input as in CAMAs [4]. These extensions will **significantly broaden the app's scope and attractiveness for continued use**.

WP2: Estimating replicability

In WP2 **I will gauge the replicability of BA studies** using a procedure refined from prominent replication initiatives [5–6]. There will be three main activities. In WP2.1 I will select studies for direct replication and redact a plan for data collection, preprocessing, and analysis. Studies will be selected from the study list compiled in WP1.1 and will comprise 16 behavioral and 2 EEG (ERP) ones—a choice motivated by both prevalence in the literature and feasibility of data collection (highest for behavioral studies and higher for EEG than fMRI). Data collection will occur on site for EEG studies and crowdsourced through Prolific or similar for behavioral ones (which might be a deviation for some of them, but not one posing issues for replicability, see [7]). **Only studies reporting a significant and clearly interpretable BA will be selected** (interpretation for ERP effects will be based on [8–9]), with priority based on: completeness of information for study reproduction; ease of reproducing the study, including recruiting a population with similar characteristics (e.g., children and older adults will be excluded from behavioral studies because they do not typically use crowdsourcing services, and those who do likely have a special profile); diversity across selected studies in terms of AC domains and bilingualism measures involved.

For both behavioral and EEG studies, data collection will involve **two phases**: 1) a screening phase in which participants will complete a short survey including the items originally used to determine monolingual/bilingual status; 2) the proper study, to which participants will be invited based on a **genetic matching algorithm** [10] aimed at maximizing the similarity with the original study sample for the responses to the survey. The study will closely reproduce the original measures involved in the BA (to maximize efficiency, measures not involved in the BA in the original study will be excluded when possible), with some precautions for crowdsourced studies to ensure consistency of stimulus appearance across devices and participants' engagement (e.g., attention checks). For each study, to account for potential inflation of the original effect size due to publication bias, a large enough sample will be collected for **90% power to detect 75% of the original effect** [5].

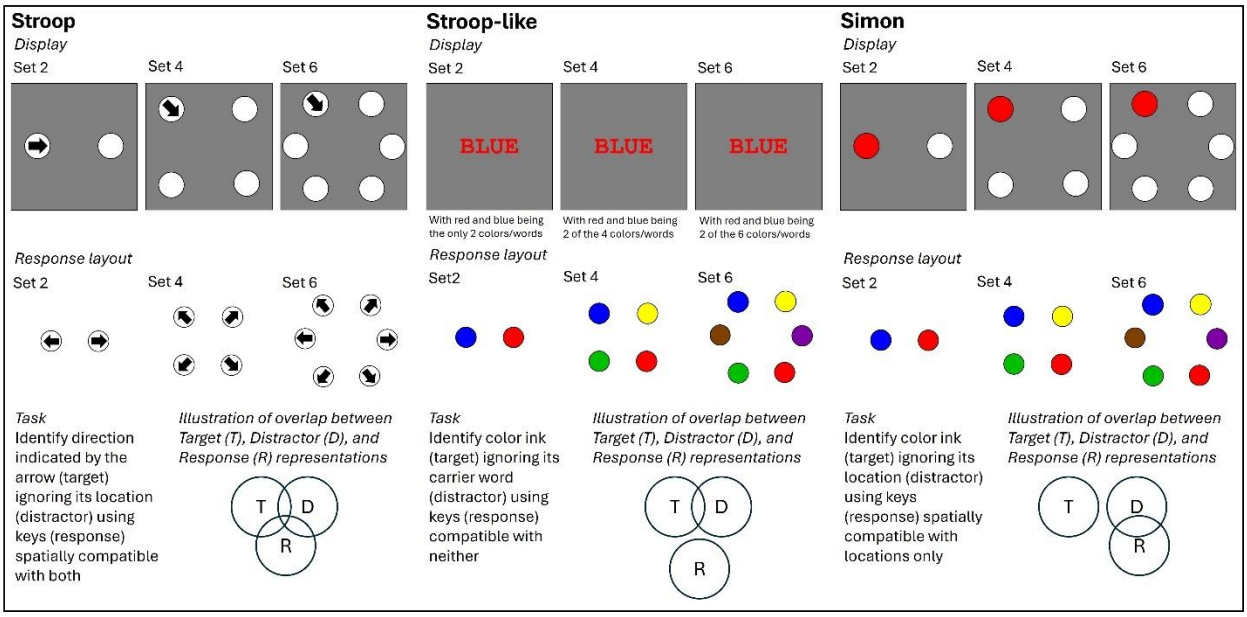
Each study will be preprocessed and analyzed with both the original procedures and all common preprocessing/analysis procedures in a **multiverse approach** to probe the robustness of the results [11]. A meta-analysis across studies will also be performed for the behavioral studies (which are in sufficient number to allow such an analysis) examining **several indicators of replication success** (e.g., statistical significance in the same direction, effect size discrepancy, replication Bayes factor) and potential moderators (e.g., domain, original study quality) [5].

A detailed **registered report** with these plans will be sent to the original authors of each study, and all possible amendments will be made to implement their advice. The registered report will then be submitted to a **top journal in the field** for Stage 1 review. Finally, in WP2.2 and WP2.3 I will carry out all final plans for the behavioral and EEG studies, respectively, once Stage-1 acceptance is

obtained. This way, even if replication rate turned out to be low, the research's validity would not be questioned.

WP3: Enhancing replicability

In WP3 I will improve the replicability of existing BA studies by addressing **all of its known threats** in two main activities corresponding to two lab studies. In WP3.1, I will focus on the threats to AC tasks (leaving momentarily bilingualism in the background) in a behavioral study that, to address **construct validity,** will involve a **cohesive set** of three tasks, theoretically defined by the relationship between their targets (Ts), distractors (Ds), and responses (Rs) (Stroop: T–D–R overlap; Stroop-like: T–D overlap; Simon: D–R overlap); to address **reliability**, each task will involve **~4000 total trials** split evenly among congruent and incongruent conditions; and to address **insufficient difficulty**, each task will come with a set size (comprising Ts, Ds, and Rs) of either 2 (baseline), 4, or 6, a **simple yet effective** difficulty manipulation. The tasks are illustrated in **Figure 1**.



Data will be collected on site from 24 young adults (8 per set size, for a **~288000 observations total** exceeding that of a powerful similar study [12]) over eight 2-hour sessions within a 4- to 6-week period. I will document the full pattern of results using **state of the art analyses**, with a focus on determining which of the larger (i.e., harder) set sizes produces the **best compromise in terms of reliability within tasks and correlations across tasks** at reduced trial sizes.

Task versions with that set size will be used for the EEG study in WP3.2. The trial size, based on [12], can be estimated at ~1000 trials per task. In a **bold combination of small-N and individual-differences research**, this still large trial size will be presented to 80 young adults, for a **~80000 observations per task** (~240000 observations total) that not only surpasses reference EEG studies (e.g., [13]), but also has 90% power to detect an individual-differences effect (e.g., a BA effect) as small as $f^2 = 0.0002$ (simulated based on [14]). A two-phase participant recruitment procedure similar to that described for WP2 will be used to maximize variability along dimensions commonly associated with a BA (while avoiding confounds) and **further boost the chances of detecting such an effect**.

There will be three sessions within a 2- to 3-week period. Sessions 1 and 2 will involve EEG recording during AC task performance. Session 3 will comprise a battery of measures including a language questionnaire and a proficiency test for subjective and objective language measures, respectively. Finally, for the next 10 days, participants will complete an EMA involving an adapted language questionnaire presented at regular intervals on their phones 4 times a day.

A comprehensive analysis set will first ascertain individual measures' psychometric properties (including for the EEG signal, with preprocessing being guided by recommended **data quality checks** [15]) and then examine associations within and between bilingualism and AC ones, considered both individually and at the latent variable level. EEG analyses will focus on AC-associated **ERP components** (e.g., lateralized readiness potential, error-related negativity) but will also examine AC-associated **frequency bands** (e.g., theta [16]). **In OPEN BAR's full spirit**, the study will be preregistered and all information/resources to reproduce it will be publicly shared.

References

1. Wilkinson et al. (2016). https://doi.org/10.1038/sdata.2016.18
2. Schneider et al. (2022). https://doi.org/10.1177/09567976221097499
3. Röseler et al. (2024). https://doi.org/10.5334/jopd.101
4. Tsuji et al. (2014). https://doi.org/10.1177/1745691614552498
5. Camerer et al. (2018). https://doi.org/10.1038/s41562-018-0399-z
6. Open Science Collaboration (2015). https://doi.org/10.1126/science.aac4716
7. Zwaan et al. (2018). https://doi.org/10.3758/s13423-017-1348-y
8. Antoniou (2023). https://doi.org/10.3758/s13423-023-02245-x
9. Cespón & Carreiras (2020). https://doi.org/10.1016/j.neubiorev.2020.07.030
10. Lukasic et al. (2018). https://doi.org/10.1371/journal.pone.0205916
11. Steegen et al. (2016). https://doi.org/10.1177/1745691616658637
12. Lee et al. (2023). https://doi.org/10.31234/osf.io/rj2bu
13. Kappenman et al. (2021). https://doi.org/10.1016/j.neuroimage.2020.117465
14. Spinelli et al. (2022). https://doi.org/10.1037/xlm0001095
15. Zhang et al. (2023). https://doi.org/10.1111/psyp.14530
16. Cavanagh & Frank (2014). https://doi.org/10.1016/j.tics.2014.04.012

Expected impact

OPEN BAR's impact will be no short of **groundbreaking**. For more than a decade, the BA hypothesis has been allegedly in a crisis, yet it **keeps attracting researchers' and funders' resources**—and understandably so, given that the hypothesized advantage would not only optimize normal cognitive function but also give it a head start in children and preserve it for longer in older adults [1]. Indeed, if the BA hypothesis were correct, bilingualism would be **one of the very few examples of cognitive training enabling far transfer** [2] and would thus demand paramount attention in **policy making**, from early childhood education to healthy aging programs. Still, since the beginning of the crisis, the main outcome of BA research appears to have been the creation of a general sense of uncertainty.

By enabling and estimating BA replicability, OPEN BAR will **dispel much of this uncertainty**, allowing a clearer picture to emerge and researchers, funders, and policy makers to make more

informed decisions. By enhancing replicability, it will also **pave the way for more robust protocols**, allowing more efficient use of available resources. As a result, OPEN BAR will allow a determination of which combinations of AC and bilingualism measures are more likely to be the locus of a replicable BA. This will have cascading effects both on the theoretical side, with the development of more refined models (addressing concerns that current BA models are too vague to be of any real use [3]), and on the practical side, with the **redirection of human, financial, and societal efforts towards the most promising areas**.

At a higher level, OPEN BAR will also stand as a **banner to rally diverse factions towards a common cause**. While this sort of unification has been repeatedly called for [3–4], international collaboration remains rare in the BA literature, in stark contrast to the general direction science is taking. OPEN BAR will address this critical shortcoming by providing **purpose and support to worldwide collaboration** on BA research. The OPEN BAR app, in particular, will not only provide an initial batch of information for replication purposes—it will also allow its users to augment that batch with both published and ongoing studies, at the same time **increasing the visibility of BA studies**, now reunited in one place, and speeding up their communication in a standardized, concise format. The evolution of the app into a more general tool for sample planning and meta-analysis with self-augmenting capabilities will further support its transition **from a small-team endeavor to a community-driven project.**

Indeed, this type of app might make a **broader contribution to open science**, which currently appears to have no middle option between repositories curated by individual authors and generalist replication hubs. Research, however, would greatly benefit from that sort of option, revolving as it does around common themes for which distinct languages, ways of reasoning, and research protocols have been developed, and that would naturally require dedicated tools. The OPEN BAR app will serve as **a model for such a tool in other areas and disciplines**. More generally, OPEN BAR will represent a shining example of open science, with a real opportunity to elevate the visibility of such initiatives (e.g., the Italian Reproducibility Network and similar European networks).

These outcomes will be favored by several **dissemination activities** aimed at raising awareness for OPEN BAR in the scientific community. Progress on the work will first be shared at the **main international conferences** of relevance, for which I plan to organize at least one symposium or satellite event. The project's primary products (in the OPEN BAR app and the behavioral/EEG replications and original studies) will also be showcased in at least **5 publications in top journals of the field** and in a **final workshop** organized towards the end of the project. Replication outcomes will also be added to the FORRT replication database [5].

Not only will these activities help OPEN BAR's visibility—they will also represent precious opportunities to start international collaborations spanning **beyond OPEN BAR's scope and duration**. To further boost such opportunities, I plan on applying to several initiatives dedicated to open and big-teams science, such as **OSCARS** (https://oscars-project.eu/) and the **Psychological Science Accelerator** [6], to fund/sponsor an international version of OPEN BAR's WP2 in which data collection will not be centralized but distributed across many teams having access to diverse bilingual populations. In sum, rather than just sealing a debate, with OPEN BAR I aim to **kickstart a long, cooperative line of research** leading the way in advancing our understanding of BA effects.

Last but not least, a replication crisis goes beyond mere replication failure—it also **damages the trust among researchers and the field's public image** [7]. The BA crisis makes no exception. Direct replication has been key in such situations, at one time exposing the reality of the problem and allowing science to re-establish its validity in the academic and public eye [8]. With its focus on direct replication, OPEN BAR will naturally favor this self-correction process and, ultimately, a **restoration of BA research's trustworthiness**.

References

1. Bialystok & Craik (2022). https://doi.org/10.3758/s13423-022-02057-5
2. Paap et al. (2024). https://doi.org/10.1177/13670069231214599
3. De Bruin et al. (2021). https://doi.org/10.1162/nol_a_00038
4. Leivada et al. (2020). https://doi.org/10.1017/S1366728920000358
5. Röseler et al. (2024). https://doi.org/10.5334/jopd.101
6. Moshontz et al. (2018). https://doi.org/10.1177/2515245918797607
7. Anvari & Lakens (2018). https://doi.org/10.1080/23743603.2019.1684822
8. Nosek et al. (2022). https://doi.org/10.1146/annurev-psych-020821-114157