EUROSLA KEYNOTE

Policy recommendations for language learning: Linguists' contributions between scholarly debates and pseudoscience

Raphael Berthele

Some language-acquisition researchers not only pursue their scholarly agenda but also act outside academia as experts in language policy-making. However, the relationship between scholarly quality and political impact is complicated, and oftentimes policy is not based on robust scholarly evidence. In this contribution, I focus on research findings in language learning that have been taken up in language planning and policy (e.g., the notion of linguistic interdependence). Drawing on concrete cases, I discuss problems of individual expertise and quality of research. Where there are methodological inadequacies and/or lack of expertise, problematic or even utterly false conclusions can be drawn from research. A critical review of influential claims in the field of applied linguistics with respect to robustness of the evidence and its fit to the actual policy problem should allow us to determine which theories and research strands may be useful for language-policy recommendations and which are probably not. A critical review of linguists' involvement in policy-making suggests that often a more appropriate appellation for so-called evidence-based policy would be policy-based evidence. In my discussion, I address two delimitation problems: defining the boundary between pseudoscience and real science (in the wide sense of the term, including social sciences and humanities) and defining the boundary between scholarly rigor and political advocacy by academics.

Keywords: expert opinion; language policy; multilingualism; research biases

1. Introduction

Research on language learning is considered highly socially relevant, most obviously in educational and political contexts where specific types of linguistic and cultural heterogeneity (e.g., immigration) are considered a problem. It is therefore alluring to emphasize the immediate social relevance of the investigation of second-language (L2) and foreign-language (FL) learning.[1] However, this emphasis can be problematic for two reasons. The first relates to the robustness of our research: What exactly are the generalizable claims that our current theories and findings license? Do the results we produce yield clear support for our theories? The second reason relates to the fit of our findings to the policy problems that await solutions: Can we generalize our findings to the language users in the scope of the policy? Is the knowledge gathered in language-acquisition studies of the type that authorizes the formulation of recommendations?

In this contribution, I will draw on my experience as director of an institute whose goal it is to produce research that can potentially inform language-policy makers.[2] I illustrate my considerations with selected issues related to multiple language learning as examples of the general problem, deferring a more comprehensive discussion of the problem to future publications.

Policy-makers need to know how to identify experts they can trust (Willingham, 2012). The most dangerous experts are those who are unaware of their own ignorance (Kruger & Dunning, 1999). The following example in (1) is an illustration of a scholar making claims while ignoring his own ignorance, when answering the question: Could teaching multiple languages to primary school children overburden the learners?

> (1) "You only have to look at the African example to prove the opposite. There, it is not rare to see children growing up with four or five languages, and that does not pose any problems", asserts the professor. (La Liberté, 22.9.2006, translation by the author)

I am the professor in question, and the answer I gave in 2006 was ignorant (as I will spell out in this contribution), but of course I did not realize this at the time. The example illustrates thus the Dunning-Kruger effect (see also Berthele, 2018 [https://plurilinguisme.shinyapps.

University of Fribourg/Freiburg, CH

raphael.berthele@unifr.ch

io/expert-on-stuff/] for eight questions to think about when reflecting one's own expertise). At the time of this interview, I had just been appointed professor in multilingualism at the University of Fribourg; I had some experience in research on transfer in multiple-language learning and using and some knowledge of the – often programmatic rather than empirical – literature on bi- and multilingual education, but most importantly I was enthusiastic about multiple-language learning and my new job in this field. However, as I argue in the present contribution, my answer was based on surrogate outcomes, on some superficial knowledge of rather irrelevant contexts (multilingual societies with low literacy rates versus the Western educational systems) as well as on evidence of questionable robustness on bilingual education. Moreover, I was surrounded by people holding similar beliefs (i.e., group conformity bias), and our beliefs tied in nicely with our political ideals. Furthermore, I was aware that my position had been created as a response to important changes and challenges related to multilingual education. In this context, my institution expected me and my colleagues to acquire external funding for multilingualism research, funding that was and still is associated with a multilingual policy agenda (I call this the *minstrel problem*). Thus, in contrast to the general complaint of applied linguists I hear in my context ("Why don't politicians listen to us experts?"), I will be concerned in this contribution with self-criticism by asking two questions:

- Am I sure I am an expert in the matter?
- Is there sufficient scholarly evidence to give recommendations?

My goal is not to attack socially engaged applied linguists but to identify some problems that, if they remain unaddressed, can make our work irrelevant both as regards the advancement of our discipline and as a basis for policy recommendations.

2. How can we tell good from bad science in applied linguistics?

2.1. Pseudoscience: Delimitation problem

The delimitation of science and pseudoscience is the topic of book-length publications (cf. Pigliucci & Boudry, 2013, for a recent example). Applied linguistics and the study of L2 acquisition are sometimes referred to as *soft science* – they may not be (hard) *science* in the narrow sense of the term, but they should also not be *pseudoscience* (Mahner, 2013, p. 25). Occasionally, scholars in our field take a Popperian stance (Hulstijn, 2015, p. 26), which relies on one main criterion to demarcate true scholarship from pseudoscience: scientific theories should be falsifiable. Others apply less strict criteria (e.g., converging evidence from multiple studies as a basis for the confirmation of theories (Mahner, p. 26)). Research on L1, L2, and L3 acquisition deals with phenomena that depend on a multitude of factors that are often difficult to measure. For research in our field to yield results that can inform educational policies, I suggest the following three assumptions as minimal requirements: (1) central

tenets of the main theories in our discipline are amenable to measurement, (2) the theories that deserve attention yield hypotheses that can be tested, and (3) it is possible to make generalizable claims about language learning that allow prognosis and not just retrodiction.

I am aware that these three points are not uncontroversial. There are several approaches to L2 and L3 acquisition that I deem to be scientific but that do not fulfil these three criteria, which is why they lie outside the scope of the present contribution. As an example, qualitative approaches that seek to shed light on the mechanisms of interactions in selected tokens of discourse (Pekarek Doehler, 2010) certainly question requirements (1) and (2) and therefore probably also requirement (3). Another example would be dynamic systems theory (DST) based approaches. They are often characterized as meta-theoretical frameworks that "unify" (de Bot et al., 2013, p. 216) mid-level theories and therefore do not directly yield testable hypotheses. Moreover, some scholars working within DST approaches focus on retrodictions rather than predictions (Larsen-Freeman, 2016, p. 388), which is due to the complex and hard to predict nature of the language-acquisition process in this view. While I have great respect for in-depth qualitative analyses and in the investigation of individual and intra-individual variability, I believe that it is also worth trying to make generalizable claims about language learning. Such generalizable claims, however, require evidence from high-quality quantitative research. At the risk of repeating myself, I do not claim that qualitative research or case studies are unscientific, but that policy recommendations need to be based on a different type of scientific evidence (requirements 1–3 above). All that follows here is based on this assumption.

2.2. Definitions

I suggest three rough working definitions categorizing research and research-like activities in our field. *Science* involves constructing theories explaining puzzling phenomena and empirically testing hypotheses derived from these theories. Scientific inquiry does not have to start from theory, it can also involve bottom-up observation and exploration of phenomena, which later allows for theory construction and revision. Scientific knowledge is unstable because it often advances by proving the old theories wrong and proposing new, better theories. Scientists change their beliefs if there is enough evidence to do so. What *enough evidence* means is controversial, especially in humanities. There is good science and bad science, but both are science and are amenable to critical assessment by the scientific community. However, if a theory is compatible with all potential data patterns in empirical investigation, it belongs to the realm of pseudoscience. Pseudoscience is based on theories that are vague, that yield untestable predictions or predictions that always apply. The output looks like science, but it does not contribute to the advancement for the field since theories, beliefs and assumptions cannot be proven wrong. Both bad science and pseudoscience need to be distinguished from what Frankfurt (2005) calls *bullshit*. Bullshit refers to texts produced by people who want to impress without caring about facts, whereas authors of bad science and pseudoscience indeed do care

about facts and truth. Indeed, in the case of pseudoscience they often care too much (see below).

2.3. What does "Research has shown…" in policy discourse refer to?

If a policy recommendation bears the badge "research has shown…", the research in question might simply not exist, the recommendation might be based on misinterpretation of solid research or on pseudoscientific beliefs, on bad science, but it is also possible that there are indeed robust findings underpinning the policy recommendation.

A policy recommendation in the domain of language teaching and learning that claims to be evidence-based typically involves the following elements:

· A curriculum-related measure Z (e.g., teach HLs to children of immigrants, start teaching two FLs on the primary school level)
· Reference to research that has shown a causal effect of one entity on another (X→Y in **Figure 1**; e.g., early age of onset → better proficiency; L1 → positive transfer to L2; first FL → positive transfer to second FL)
· The assumption that Z allows unfolding the effect of X on Y (i.e., Z→(X→Y))

The terms *programmatic literature* and *doxa* in **Figure 1** might need some explanation. As will be shown in two examples in the next section, there are cases in which the efficacy of a measure Z only relies on literature that is programmatic (e.g., literature that makes claims on Z, possibly based on evidence on X→Y, without providing any evidence on the effects of Z). If groups of actors in the scholarly field take such programmatic claims and their entailments for granted, then this shared common ground represents what Bourdieu (1979, p. 549) refers to as the doxa.

As **Figure 1** shows, there are many cases where no recommendation should be made. Three configurations, however, license recommendations with increasing certainty. In **Figure 1** I do not address the possibility that a policy measure Z has beneficial effects even though the evidence as to X→Z is inconclusive. Such cases undoubtedly exist, just as certain substances have proven to have a therapeutic effect although research does not fully understand why. The focus of my contribution, however, lies in the possibility for second language acquisition (SLA) researchers to make evidence-based recommendations, for which having converging evidence of X→Y is a prerequisite.

3. Case study I: Recommending HL classes

The first case I would like to discuss is the recommendation formulated in (2).

(2) There is evidence supported by practition-ers that the following contribute to raising the attainment of children without the language of instruction: […] Developing their mother tongue competences. (European Commission, 2015, p. 5)

The underlying rationale is that migrant children should attend HL classes (Z) because they foster linguistic development in HLs that in turn benefits the learning of the L2 (X→Y). This view of interdependent language skills is deemed particularly relevant for cognitively demanding areas of language use and proficiency (cf. Cummins, 1996). The argument is important in many contexts where some variant of bilingual education involving HL and L2 is advocated: It is usually directed against criticism from
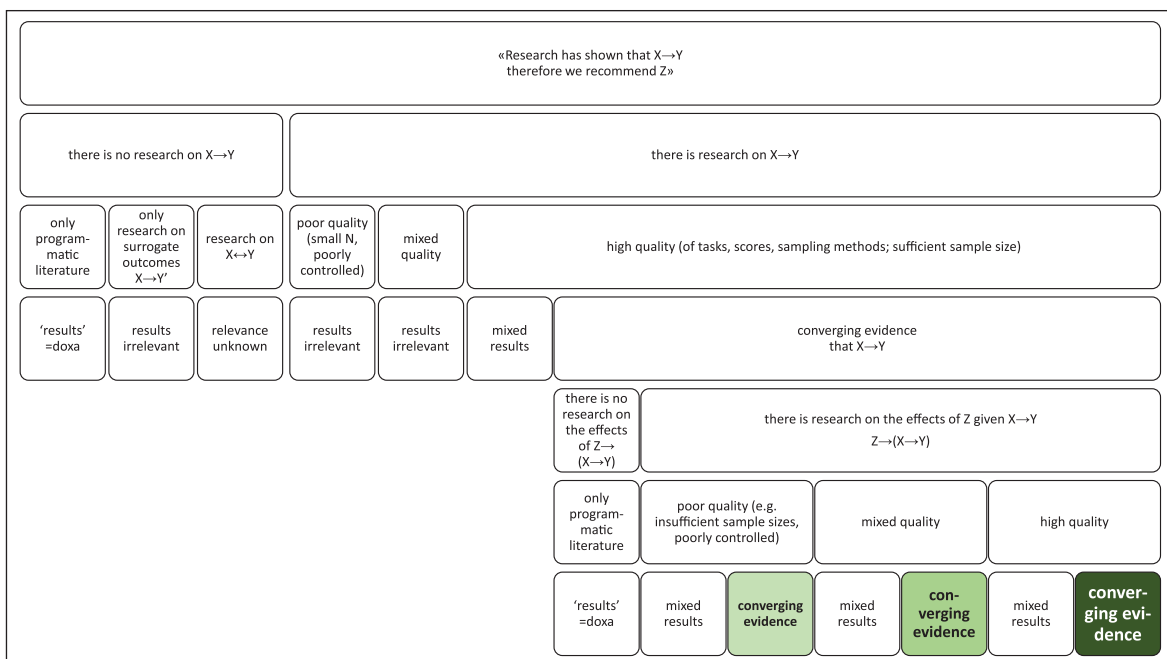

Figure 1: Research and policy recommendations. Evidence-based recommendations should only be made in the cases marked in bold/green. Directed arrows ('→') refer to causation, bidirectional arrows ('↔') to association without entailing causation.

approaches that emphasize the importance of exposure time for the learning/acquisition of a specific language without assuming noticeable transfer effects (time-on-task, see discussion below).

The interdependence assumptions are appealing to many of us for several reasons. They tie in with the holistic view on bi- and multilingual repertoires (Cook, 1995), and they value transfer (Odlin, 1989).[3] Such theories also provide a rationale for more language learning in general and for language rights of a vulnerable group of learners. In the following sections, I briefly review the literature on the scholarly evidence for interdependence. This will allow us to determine whether the recommendation is based on sufficient evidence and hence is an instance located on the right-hand side of **Figure 1**.

A comprehensive discussion of the evidence on X→Y would be a book-length enterprise and cannot be done here (see Berthele & Lambelet, 2017b, for an overview). I will therefore limit my discussion to selected aspects that I consider prototypical for the central issue of my contribution.

3.1. Lack of fit: Surrogate outcomes

When it comes to the effects of L1s on L2s (X→Y), there are some robust findings. There is ample evidence showing all kinds of cross-linguistic (bi- and multi-directional) influence on many different levels, from phonology to syntax and pragmatics (e.g., Cenoz, Hufeisen, & Jessner, 2001). But this evidence mostly pertains to fundamental research, not to teaching-oriented research. Moreover, this research includes evidence about what is often termed negative transfer. Evidence from cross-linguistic influence in SLA and bilingualism research is not a fit basis for the policy recommendations. It is almost never based on intervention studies, but on studies of bi- and multilingual subjects, and these studies often involve rather controlled tasks: Psycholinguistic experiments or picture descriptions with very specific stimuli, to mention just two research paradigms, are tasks that are only indirectly related to naturalistic (school) language tasks. Evidence from such studies is scientifically relevant and can contribute to theory development, but with respect to pedagogical practices or policy recommendations they represent *surrogate outcomes* (Y' in **Figure 1**). They have limited to zero relevance for an argument about literacy acquisition in multilinguals: The relationship between the psycholinguistic investigation of lexical access in bilinguals and policy recommendations regarding HL instruction is about as tenuous as the relationship between testing the effects of a molecule on cancer cells in a petri dish compared to the effect of a cancer therapy on survival rates of cancer patients (see Sutherland, 2013, for other examples of the surrogate outcome fallacy). Robust findings from language-use patterns as those practiced in school settings are needed.

3.2. Lack of robustness: Methodological shortcomings

Often, the quality of research into interdependence is mixed. From the point of view of theory testing, the gold standard would be research that randomly assigns learners to an experimental group where L1 literacy is taught effectively and to a control group where some other cognitively interesting activity not related to the L1s is carried out. The effects of such an intervention then would be measured over a period of several years since learning literacy takes several years. Such research is not available, for obvious reasons: It would be highly problematic to randomly select young learners and force them either to learn their HL or inversely to prohibit this learning for others. Thus, research on HL learning is usually based on survey data of self-selected samples of HL earners (cross-sectional or longitudinal).

Many scholars find positive linear associations between skills in L1s and L2s (see Berthele & Lambelet, 2017b, for an overview), and such findings are then interpreted as evidence for interdependence, especially if the learners with HL instruction outperform the learners without HL instruction. Given that the groups getting HL instruction are overwhelmingly self-selected samples, such studies are biased towards confirmation of the interdependence theory: The phenomenon is studied "in winners" (Goldacre, 2009, chapter 11). What is interpreted as a positive effect of HL instruction (on HL or on L2) is at least in part the effect of a sample where the not-so-linguistically gifted are underrepresented. Ignoring this leads to naïve statements (see example 1).

3.3. Theoretically expected direction of transfer effects

As studies often did not yield the expected positive associations, initial, simpler versions of the theory were reformulated as in (3).

> (3) To the extent that instruction in Lx is effective in promoting proficiency in Lx, transfer of this proficiency to Ly will occur provided there is adequate exposure to Ly (either in school or environment) and adequate motivation to learn Ly. (Cummins, 1996, p. 111)

In even later developments, many scholars assume that there is two-way transfer across languages. Moreover, as discussed in Hulstijn (2015, p. 117), several variants of threshold theories are added to the idea of positive transfer in literacy skills, sometimes involving minimal levels in the L1, sometimes involving minimal levels in the L2 (e.g., Alderson, 1992). Such thresholds are hardly ever established a priori (i.e., before the data are analysed) but developed a posteriori: Once differences in the strength of linear cross-linguistic associations of skills are found, the threshold explanation is used to explain the pattern (see discussion below and Takakuwa, 2005, for a critical analysis of threshold theories). Cummins (1979, p. 230) argues that thresholds vary within and across individuals, which poses similar problems for policy recommendations as the case study perspective of DST discussed above.

Modifying the interdependence theory by adding elements such as unspecific thresholds or vaguely spelled out contextual conditions required for successful learning can lead to problematic uses of empirical findings: Explaining results post-hoc invoking such elements can

be considered HARKING (hypothesizing after the results are known, Kerr, 1998). Instead of testing a hypothesis using data, we sometimes use data to find hypotheses which then are unsurprisingly consistent with these same data. Of course, this is by no means to say that we should never revise our theories based on unexpected patterns in our data. In the context of threshold theories, as an example, an interesting way to go could be to infer possible thresholds from patterns in the data and then put these thresholds to the test with new data.

Experimental evidence for the influence of learning one language on learning another is scarce, for the reasons spelled out above. In our longitudinal project on Portuguese as a heritage language, we therefore tried to come up with predictions based on the interdependence idea that are testable. We gathered data from a sample of Portuguese children in both French- and German-speaking Switzerland. One of our predictions was that literacy, which is developed in the L2 in a much more intensive way than in the HL, would be transferred to the HL in a cross-lagged pattern. That same optimistic take on inter-lingual transfer of skills had shaped my view expressed in the newspaper interview (1). However, our data show that although skills measures across languages and within languages were all positively correlated, there is no statistically significant larger effect from L2 to HL than in the other direction (for details on these analyses see Berthele & Vanhove, 2017).

There are several possible patterns in correlational data gathered in the HL and SL context:

    A. a negative linear association of HL and L2 skills;
    B. no association between HL and L2 skills;
    C. a positive linear association of HL and L2 skills;
    D. non-linear associations (modifying patterns A or C, e.g., due to thresholds)

In short, a multitude of data patterns are compatible with the interdependence plus threshold theory: If there is no association or if it is negative, then scholars can argue that the necessary threshold has not been reached in one or both languages or that the learners lacked motivation to learn the HL. If the association is non-linear, then this can be interpreted as evidence for thresholds. If it is positive and linear, this can be considered as evidence of interdependence in a more straightforward way. However, positive correlations must not be confounded with causality since they can be due to all kinds of other effects, ranging from socio-economic correlates to task-wiseness to general cognitive abilities that are not related to language exclusively. Many influential scholars in the field do not advocate a strict separation between linguistic skills and other cognitive attributes. This is certainly true for most of the usage-based approaches to language learning (e.g., Wulff & Ellis, 2018) but also for interdependence-related research (e.g., "transfer and [cognitive] attributes are two sides of the same coin." Cummins, 2017, p. 107). Such a fused construal of linguistic and cognitive skills, in my view, is theoretically and empirically sound. However, it raises important questions regarding the policy recommendation that is at stake: If the common

proficiency underlying the claim that investments in the HL are good for the L2 is (largely?) formed by general cognitive skills (memory, fluid intelligence, etc.), then it becomes rather unclear why investing in the learning of one language should be particularly beneficial for another language. To the extent that these general cognitive skills can be trained and transferred, any cognitively interesting activity should therefore be shown to have beneficial effects on literacy acquisition in the L2.

In sum, as others have argued before (e.g., Hulstijn, 2015, p. 130), we are dealing with a theory that has many flaws. In my view its main problem is that it intuitively makes sense while being too powerful: It is true regardless of what the data say. A theory that is consistent with any imaginable type of finding pertains to pseudoscience, resembling astrology (Pigliucci & Boudry, 2013, p. 16, Figure 1.1). More importantly, unless the thresholds or the contextual conditions and proficiency levels required for transfer are specified, no recommendations should be given to policy-makers, since the effect of X→Y is one among several possible consequences of HL instruction (Z) occurring under certain ill-specified conditions.

3.4. Interdependence: A discursively successful scholarly failure

Along the lines laid out in **Figure 1**, we can thus conclude that the evidence for X→Y is inconclusive, although many scholars find significant positive associations of skills across the two languages.

The methodological shortcomings of the research make it impossible to ascertain causal influences from HL literacy skills to L2s (or vice versa, for that matter). The interdependence approach, if supplemented post hoc by threshold explanations, is vague and does not lend itself to hypothesis testing and to the advancement of theory. Given this situation, research on Z (i.e., on the impact of HL instruction on X (which in turn influences Y)) is superfluous if its goal is to put the whole theory Z→(X→Y) to the test. In our recent book (Berthele & Lambelet, 2017a), several studies on HL and L2 development were presented, and the methodologically most compelling is a study by Moser et al. (2017). This quasi-experimental, longitudinal study potentially allows statements about Z→(X→Y) because it compares skills from a control group and an experimental group, and only this experimental group had followed a carefully designed pedagogical intervention in the respective HLs. The study, however, did not even find an effect of HL instruction on HL proficiency. Of course, this is only one study, but other intervention studies that test the effects of pedagogical measures fostering bilinguals' languages find similar results (e.g., McElvany et al., 2017). As is often the case, the higher the quality of the empirical investigation, the smaller the effects of the supposed beneficial measure.

To sum up, the evidence for causal transfer effects from L1 to L2 in the literacy domain is inconclusive (the effects are spurious, the causality behind the linear association is unclear), and the evidence for positive effects of HL instruction on L2 is scarce. Recommendations of HL instruction based on this evidence, with the rationale that proficiency in HL will foster proficiency in L2, thus

seem to lack an adequate empirical base. The notion of linguistic interdependence is above all used to argue in favor of bilingual education, in contexts where speakers of particular HLs form a large proportion of the pupil population. Of course, bilingual education and various forms of HL instruction can be considered necessary or good for other reasons, and therefore research on HL instruction may still be relevant (I will come back to this in the concluding section). But ideas of transfer-benefits based on the interdependence theory are currently not supported by robust evidence.

4. Case study II: 'Early' instruction of two FLs

The second case study does not need very much introduction in the context of SLA studies. Like many other European countries, Switzerland has already introduced so-called early FL instruction in primary school. Quote (5) is from a policy document that outlines the main goals and strategies of FL instruction in Switzerland.

> (5) [T]he pupils will develop proficiency in at least one other national language; [...] c) the pupils will develop proficiency in English; [...] e) pupils speaking a foreign native language will get the opportunity to consolidate this proficiency. [...] This important goal can only be reached if:
>
> - the teaching and learning of languages in general is improved sustainably
> - the potential lying in early language learning is used optimally, which means introducing successively two foreign languages by 5th grade at the latest. (EDK, 2004, p. 2, translation by the author)

This strategy was adopted by the Swiss cantonal ministers of education in 2004. The Swiss primary school curriculum thus involves in almost all cantons the onset of two FLs before secondary school (starting at about age 12). However, the minimal standards to be achieved by students in all tracks in both languages are the same, despite a two-year difference between the onsets of the L1 and L2. The policy-makers appeal to transfer processes from the first to the second FL, which is reminiscent of the transfer versus time-on-task discussion in the previous section on interdependence:

> (6) In 11th grade the basic competence levels are identical for the two foreign languages. The comparatively faster learning progression in the second foreign language is possible because of the benefit due to the competences learned in the first foreign language. (CDIP, 2011, p. 7 translation by the author)

At the time of this curriculum reform, there was already quite substantial research on the effects of age of onset (AoO) both in L2 and FL learning (see Lambelet & Berthele, 2015, for an overview).

Whereas scholars in our field disagree about the critical period hypothesis both from a theoretical point of view and from the point of view of the available evidence,

this discussion is almost exclusively concerned with the acquisition of a L2 (i.e., a language spoken in the learners' environment), as is typically the case in migration contexts. It is important here to make the distinction between L2 and FL learning, the latter referring to the learning or the acquisition of a FL in a classroom setting with limited exposure (typically 2–4 weekly lessons). And even though we may fundamentally disagree about the age factor and critical periods in L2 acquisition, there is agreement that an earlier start of FL teaching does not consistently lead to better proficiency. In the Swiss context, there was a heated debate because of popular votes on initiatives (a Swiss variant of a referendum) that aimed to abolish the teaching of two FLs at primary school level. Moreover, there is also debate about the priority that should be given to English or a national language as the first FL (see also Mittler, 1998; Ronan, 2016). Thus, once again, we need to distinguish between political reasons for teaching specific languages in the curriculum and scholarly evidence concerning the effects of earlier versus later AoO.

4.1. Lack of fit: Surrogate outcomes

As mentioned above, the fundamental problem with the debate on AoO in FL instruction is that it mainly draws on evidence from L2 acquisition. Most colleagues, regardless of their theoretical affiliation, would agree that there are age effects if one examines cohorts of L2 learners, and that these effects, regarding proficiency in the long run, can be grossly summarized as 'the younger the better'. However, such studies focus on surrogate outcomes when FL rather than L2 learning is at stake. In the Swiss case as in many others, discussion also drew on neurolinguistic evidence concerning the activation patterns of younger versus later bilinguals, with an earlier AoO yielding more overlapping activation in specific areas of the brain than later AoO (see, for example, Imgrund & Le Pape Racine, 2005). The study often cited in this debate, however, did not focus on proficiency but on age-related differences in the locus of brain activity; proficiency in earlier and later learned languages was held stable (Wattendorf, 2006).[4] The main researcher of the study, the neurologist Elise Wattendorf, never drew any curriculum-oriented inferences regarding what the different age-related activation patterns mean (see Berthele, 2014, for a discussion). These brain-activation patterns, interesting as they may be for brain researchers, are not even a surrogate outcome here and cannot serve as a basis for curriculum planning. Instead of critically reviewing the research on the effects of AoO in FL teaching, many scholar-advocates enriched their publications by irrelevant references to brain research, which brings to mind McCabe and Castel's study (2008) showing that research reports including irrelevant information on brain research are considered more convincing by readers.

What is needed to decide whether an earlier start of FL teaching is beneficial or not is again experimental, or at least quasi-experimental, research that compares maximally comparable cohorts of learners that are different with respect to AoO.

As to the idea of transfer from the first to the second FL, there is indeed evidence from bi- and multilingualism research that learners transfer in many different

directions (Pavlenko & Jarvis, 2002). Although no explicit references are given, the recommendation in (5) is most likely influenced by holistic and transfer-oriented theories of the multilingual repertoire, as is a large body of the programmatic literature on multilingual language learning (see the discussion in section 3 and the chapters in Hufeisen & Neuner, 2004, for further illustration). From our own studies on receptive skills in multilinguals, I am aware of a long-standing interest in spontaneous positive transfer of one language in the repertoire to another (see Vanhove & Berthele, 2015, for references). Studies from bilingual contexts yield mixed results: Sometimes bilinguals outperform monolinguals in L3 learning and sometimes they do not (cf. Cenoz, 2003, for an overview). However, for the Swiss curriculum reform in question, such studies are again surrogate outcomes in the sense that the proficiency acquired after two years of 2–4 weekly lessons in, say, French as a FL cannot be expected to be on a par as a basis for transfer with one of the two languages mastered by early bilinguals in Basque and Spanish in the Basque country. Also, the spontaneous positive transfer researchers in receptive multilingualism generally observe is often investigated in adults, and, even worse, in undergraduate students of linguistics or philology. Inferences from such gifted language users and learners to the population of children in compulsory primary school are risky, and it is a general feature of the programmatic literature on (early) FL learning that it is mainly informed by the introspection of language experts and by surrogate research outcomes from badly fitting samples of multilinguals.

4.2. Lack of robustness: cherry picking, CARKing and HARKing

There is one study of transfer from the first to the second FL that is often cited in the Swiss context (Haenni Hoti et al., 2011). This study, in my view, is a well-designed comparison between pupils from cantons that, at the time of data collection, just had or had not yet introduced two FLs at primary level. The study found better skills in French (as a second versus first FL) listening and reading at the end of the fifth grade. This is often taken as proof for such a beneficial transfer effect from one to the other FL. An extension of this study measured the same effects once more one year later. By then, the positive effects of French as a first FL on English as a second FL in listening and reading comprehension were gone (Heinzmann et al., 2009, p. 45ff.). This extension study, however, is unmentioned in important reviews. For example, in the systematic review by Dyssegaard et al. (2015, p. 81), only the results of the first publication 2011 are discussed.[5] Or, if the extension evidence is discussed, CARKING (criticizing after the results are known, cf. Nosek & Lakens, 2016) is practiced, in this case by the authors themselves: They criticize the quality of their own tests in the face of the null result, asserting that better tests would have yielded the expected result.[6]

Manno (2017) presents another highly relevant study, in a similar context immediately before and after the transition to two FLs in primary school. He shows no positive effects of the first FL (English) on the second FL

in reading skills (French) if the French skills are compared to the group that only learns French as a FL. In the concluding discussion, the author comes up with post hoc threshold explanations (Manno, p. 147) that could account for the null result. Had the prediction been that a specific level in the first FL needs to be attained before transfer to the second FL occurs, this prediction could have been put to the test. But testing the effect of skills in a first FL on a second FL and, in the absence of the expected result, postulating thresholds is a case of HARKING and does not contribute to the advancement of the discipline, let alone license any policy recommendations. As in the interdependence case, the theory seems to predict that skills are correlated unless they are not.

As we discuss in Lambelet and Berthele (2015), even if we could design an experimental study of AoO differences that would run for several years, there would still be the question of what we should compare: Learners of different ages after the same exposure time? Learners of the same age with different exposure time? Moreover, the question arises whether it is possible to use identical language tests for learners of different ages or who have been taught with different, age-adapted methods. Despite the lack of large-scale, robust evidence, the studies that are available of AoO and FL learning are far from clearly supporting the earlier onset in the sense of the converging evidence.

Like studies in the field of HL speakers, studies in the FL field are characterized by vague theories and an optimistic view of language teaching and learning. There is a danger that pedagogical reforms based on firm beliefs but shaky evidence will harm the vulnerable while the gifted learners are likely to learn even in the most adventurous pedagogical paradigms (cf. Berthele et al., 2017 on potential Matthew effects in multilingual language teaching; also Willingham, 2012, p. 16).

5. Discussion and Conclusion

When it comes to applied linguistics and L2 research and policy recommendations, two delimitations need to be established. The first is the one between science and pseudoscience. In this contribution, I have given examples of good science, bad science, and pseudoscience. I run the risk that my critique of some of the scholarly work will be perceived as patronizing and maybe as personal attacks on certain colleagues, but as my introductory example shows, I also consider some of my own past statements to have been dangerously ignorant. As implied in the quote by the astronomer Carl Sagan (7), there is nothing wrong in admitting that I would not issue that statement anymore today. The more I learn about transfer, the less I feel comfortable when asked to give recommendations.

(7) In science it often happens that scientists say, 'You know that's a really good argument; my position is mistaken,' and then they would actually change their minds and you never hear that old view from them again. (Sagan, 1987)

What Sagan refers to should also shape our field: The scientific method should allow us to move on, and it should prevent us from getting stuck with pedagogically

or socially optimistic but empirically untested (and often untestable) programmatic assumptions. There are three aspects to the danger I see in pseudoscience and bad science. First, we do not give evidence-based recommendations but produce far too much policy-based evidence because we ignore our own biases and our ignorance. Second, as a consequence of insufficient evidence, pedagogical innovations are susceptible to failure, which does not shed a very positive light on our discipline. If we want policy-makers to respect us, we should start by acknowledging our ignorance (and refrain from giving recommendations if they are not licensed by evidence), while doing whatever we can to produce robust results. Third, bad science, vague theories, and studies biased towards confirming our beliefs do not allow us to advance as a scholarly discipline.

The second delimitation problem pertains to our personal political, social values and our research practices: I am personally strongly in favour of HL instruction, for political reasons and for reasons of linguistic rights. I also think that people who do not want to continue speaking their HL should be respected. As a moderately patriotic Swiss citizen living in a bilingual environment and having raised two multilingual children, I am strongly in favour of teaching national languages (i.e., other languages than English) in Swiss schools, not just for pedagogical, but also for symbolic reasons. However, such personal values and preferences can easily get in the way of scholarly research: As a scholar, therefore, I need to acknowledge that my values and the available evidence do not always converge. If early French FL instruction does not produce the expected results, it is our duty to report this and to question the assumptions that lead to the policy recommendation. Distinguishing between personal values and scholarly evidence is crucial but not easy. Perhaps, scholars who are merely mediocre language learners and only moderately passionate about their theories might do better research: They would consider alternatives and look for evidence against their theories, which is more efficient for scientific progress than the quest for confirmation (see Chambers, 2017, p. 18 on why we learn virtually nothing when only looking for confirmation).

I would like to conclude more positively: There are indeed tools and procedures that allow us to improve our scientific endeavours.

Open science: If we publish not only our results but also our data, scripts, and materials so that others can replicate our studies (Marsden et al., 2018), we engage in a process that will lead to more robust evidence. This will allow scholars to discard theories based on spurious findings and identify areas where evidence truly converges.

Preregistration:[7] Specifying expectations (e.g., thresholds), materials and statistical analysis in advance is an excellent strategy for tackling the biases that infest our field. Even the most politically committed advocates of HL instruction can then test their theories and predictions, and if the ideologically expected result emerges from the data, the satisfaction provided by the results is all the greater.

Reclaim linguistics and language learning as our focus: Oftentimes, our desire to be pedagogically and socially relevant in our research is not validated by educational and sociological theory and evidence: Our affirmations are often pedagogically and sociologically naïve (Berthele, 2017). At the same time, we risk losing sight of our fundamental questions as acquisitionists. Maybe, ultimately, research on transfer (X→Y) in the domain of verb placement in German L2 acquisition will have some educational relevance (Z→(X→Y)), but first and foremost, our duty is to develop our theories in a scholarly way. Converging evidence can then inform pedagogical measures, but first we need to provide robust evidence concerning what is at the centre of our discipline (i.e., concerning X→Y).

Notes

[1] The term *second language* is often used both for the learning of a new language in a migration situation and for the learning of a FL in an instructed context. Some authors also use this term referring to a third or any other additional language. I use the term *foreign language* if the context of instructed learning of a non-local language is crucial to the argument, and I refer to *third language* (L3) if it is fundamental for my argument to distinguish between different additional languages in the multilingual repertoire. I use the term *heritage language* (HL) to refer to the first language (L1) of migrants.

[2] http://www.institut-plurilinguisme.ch/.

[3] By citing the work of these scholars, I do not imply that it should be taken for granted that they also endorse the specific policy recommendation (Z) discussed here. This also applies to all other references on scholarly evidence regarding X→Y in my contribution.

[4] Many authors in this debate referred to preliminary analyses of a subsample of the thesis in question that was fully published in 2006.

[5] I am not implying that the authors of the systematic review deliberately excluded this evidence from the extension. These results so far have not been published in an international journal. The time the reviewing and proof-reading process takes explains why the earlier results were published after the extension study, which is available for download as a report. This might explain why the latter study remained largely unnoticed. This, however, raises the question of why only the first set of results was published in an international journal.

[6] Although this is now a form of HACARKIKing (hypothesizing after the criticism after the results are known is known), I hypothesize that the authors would not have questioned their tests had the results shown the expected effect of the first on the second FL.

[7] See https://cos.io/rr/.

References

Alderson, J. C. (1992). Reading in a foreign language: A reading problem or a language problem? In J. C. Alderson & A. H. Urquhart (Eds.), *Reading in a foreign language* (pp. 122–135). London, New York: Longman.

Berthele, R. (2014). Broca in der Schulsprachenpolitik. *Babylonia*, *1*, 64–69.

Berthele, R. (2017). Sprachliche Heterogenität und Schule: Eine Kritik an der Linguistisierung sozialer Probleme. *Vierteljahresschrift Für Heilpädagogik Und Ihre Nachbargebiete*, *3*, 203–216. DOI: https://doi.org/10.2378/vhn2016.art45d

Berthele, R. (2018, September 6). Am I an expert on STUFF? Retrieved from https://plurilinguisme.shinyapps.io/expert-on-stuff/

Berthele, R., & **Lambelet, A.** (Eds.) (2017a). *Heritage and school language literacy development in migrant children: Interdependence or independence?* Clevedon, UK: Multilingual Matters. DOI: https://doi.org/10.21832/9781783099054

Berthele, R., & **Lambelet, A.** (2017b). Investigating interdependence and literacy development in heritage language speakers: Theoretical and methodological considerations. In R. Berthele & A. Lambelet (Eds.), *Heritage and school language literacy development in migrant children: Interdependence or independence?* (pp. 1–25). Clevedon, UK: Multilingual Matters. DOI: https://doi.org/10.21832/9781783099054-002

Berthele, R., Lambelet, A., & **Schedel, L. S.** (2017). Effets souhaités et effets pervers d'une didactique du plurilinguisme: l'exemple des inférences inter-langues. *Le Français Dans Le Monde*, *61*, 146–155.

Berthele, R., & **Vanhove, J.** (2017). What would disprove interdependence? Lessons learned from a study on biliteracy in Portuguese heritage language speakers in Switzerland. *International Journal of Bilingual Education and Bilingualism*, 1–17. DOI: https://doi.org/10.1080/13670050.2017.1385590

Bourdieu, P. (1979). *La Distinction: Critique sociale du jugement.* Paris: Les Editions de Minuit.

CDIP. (2011). Compétences fondamentales pour les langues étrangères. Standards nationaux de formation. Berne: Conférence suisse des directeurs cantonaux de l'instruction publique.

Cenoz, J. (2003). The additive effect of bilingualism on third language acquisition: A review. *International Journal of Bilingualism*, *7*(1), 71–87. DOI: https://doi.org/10.1177/13670069030070010501

Cenoz, J., Hufeisen, B., & **Jessner, U.** (2001). *Cross-linguistic influence in third language acquisition.* Clevedon: Multilingual Matters. DOI: https://doi.org/10.21832/9781853595509

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice.* Princeton, NJ: Princeton University Press. DOI: https://doi.org/10.1515/9781400884940

Cook, V. (1995). Multi-competence and the learning of many languages. In M. Bensoussan, I. Kreindler, & E. Aogáin (Eds.), *Multilingualism and language learning: 8, 2. Language, culture and curriculum* (pp. 93–98). Clevedon, UK: Multilingual Matters. DOI: https://doi.org/10.1080/07908319509525193

Cummins, J. (1979). Linguistic Interdependence and the Educational Development of Bilingual Children. *Review of Educational Research*, *49*, 222–251. DOI: https://doi.org/10.3102/00346543049002222

Cummins, J. (1996). *Negotiating identities: education for empowerment in a diverse society.* Los Angeles: California Association for Bilingual Education.

Cummins, J. (2017). Teaching for transfer in multilingual educational contexts. In O. García, A. Lin, & S. May (Eds.), *Bilingual and multilingual education* (3rd ed.) (pp. 103–115). New York, NY: Springer. DOI: https://doi.org/10.1007/978-3-319-02258-1_8

de Bot, K., Lowie, W., Thorne, S. L., & **Verspoor, M.** (2013). Dynamic systems theory as a comprehensive theory of second language development. In M. García Mayo, M. J. Gutiérrez Mangado, & M. Martínez-Adrián (Eds.), *Contemporary approaches to second language acquisition* (pp. 199–220). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/aals.9

Dyssegaard, C. B., Egeberg, J. de H., Sommersel, H. B., Steenberg, K., & **Vestergaard, S.** (2015). *A systematic review of the impact of multiple language teaching, prior language experience and acquisition order on students' language proficiency in primary and secondary school.* Copenhagen: Danish Clearinghouse for Educational Research. Aarhus: Danish Clearinghouse for Educational Research.

EDK. (2004). Sprachenunterricht in der obligatorischen Schule. Strategie der EDK und Arbeitsplan für die gesamtschweizerische Koordination. Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren.

European Commission. (2015). Language teaching and learning in multilingual classrooms. How schools and communities can help learners with different linguistic backgrounds strengthen their language skills in order to succeed better in school and life (Policy brief).

Frankfurt, H. G. (2005). *On bullshit.* Princeton, NJ: Princeton University Press.

Goldacre, B. (2009). *Bad science.* London: Harper Perennial.

Haenni Hoti, A. U., Heinzmann, S., Müller, M., Oliveira, M., Wicki, W., & **Werlen, E.** (2011). Introducing a second foreign language in Swiss primary schools: The effect of L2 listening and reading skills on L3 acquisition. *International Journal of Multilingualism*, *8*(2), 98–116. DOI: https://doi.org/10.1080/14790718.2010.527006

Heinzmann, S., Müller, M., Oliveira, M., Haenni Hoti, A., & **Wicki, W.** (2009). Englisch und Französisch auf der Primarstufe – Verlängerung des NFP-56-Projekts. Schlussbericht. Luzern.

Hufeisen, B., & **Neuner, G.** (Eds.) (2004). *The plurilinguism project: Tertiary language learning – German after English.* Strasbourg: Council of Europe.

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research.* Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/lllt.41

Imgrund, B., & **Le Pape Racine, C.** (2005). Ja zu zwei Fremdsprachen an der Primarschule. *Babylonia, 1,* 63–67.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196–217. DOI: https://doi.org/10.1207/s15327957pspr0203_4

Kruger, J., & **Dunning, D.** (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134. DOI: https://doi.org/10.1037//0022-3514.77.6.1121

La Liberté. (2009). Quotidien romand édité à Fribourg. Fribourg: Imprimerie St. Paul.

Lambelet, A., & **Berthele, R.** (2015). *Age and foreign language learning in school.* London: Palgrave MacMillan. DOI: https://doi.org/10.1057/9781137525901

Larsen-Freeman, D. (2016). Classroom-oriented research from a complex systems perspective. *Studies in Second Language Teaching and Learning, 6*(3), 377–393. DOI: https://doi.org/10.14746/ssllt.2016.6.3.2

Mahner, M. (2013). How to demarcate after the (alleged) demise of the demarcation problem. In M. Pigliucci & M. Boudry (Eds.), *Philosophy of pseudoscience* (pp. 29–43). Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226051826.001.0001

Manno, G. (2017). Lesekompetenz im Französisch als 1. und 2. Fremdsprache in einer mehrsprchigen Perpektive bei Primarschüler/innen am Ende der 6. Klasse. *Zeitschrift Für Interkulturellen Fremdsprachenunterricht, 1,* 139–151.

Marsden, E., Morgan-Short, K., Thompson, S., & **Abugaber, D.** (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning, 68,* 321–391. DOI: https://doi.org/10.1111/lang.12286

McCabe, D. P., & **Castel, A. D.** (2008). Seeing is believing: The effect of brain images on judgments of scientific reasoning. *Cognition, 107*(1), 343–352. DOI: https://doi.org/10.1016/j.cognition.2007.07.017

McElvany, N., Ohle, A., El-Khechen, W., Hardy, I., & **Cinar, M.** (2017). Förderung sprachlicher Kompetenzen. Das Potenzial der Familiensprache für den Wortschatzerwerb aus Texten. *Zeitschrift Für Pädagogische Psychologie, 31*(1), 13–25. DOI: https://doi.org/10.1024/1010-0652/a000189

Mittler, M. (1998). *Wieviel Englisch braucht die Schweiz? Unsere Schulen und die Not der Landessprachen.* Frauenfeld: Huber.

Moser, U., Bayer, N., & **Tomasik, M. J.** (2017). Language skill transfer effects: Moving from heritage language to school language in kindergarten. In R. Berthele & A. Lambelet (Eds.), *Heritage and school language literacy development in migrant children:*

Interdependence or independence? (pp. 119–140). Clevedon, UK: Multilingual Matters. DOI: https://doi.org/10.21832/9781783099054-008

Nosek, B. A., & **Lakens, D.** (2016). *Registered Reports: A method to increase the credibility of published reports.* PsyArXiv. DOI: https://doi.org/10.31234/osf.io/mgxyg

Odlin, T. (1989). *Language transfer: cross-linguistic influence in language learning.* Cambridge: Cambridge University Press. DOI: https://doi.org/10.1017/CBO9781139524537

Pavlenko, A., & **Jarvis, S.** (2002). Bidirectional transfer. *Applied linguistics 23,* 190–214. DOI: https://doi.org/10.1093/applin/23.2.190

Pekarek Doehler, S. (2010). Conceptual changes and methodological challenges: On language and learning from a conversation analytic perspective on SLA. In P. Seedhouse, S. Walsh, C. Jenks (Eds.), *Conceptualising 'learning' in applied linguistics* (pp. 105–126). London: Palgrave Macmillan. DOI: https://doi.org/10.1057/9780230289772_7

Pigliucci, M., & **Boudry, M.** (Eds.) (2013). Philosophy of Pseudoscience. Chicago: University of Chicago Press. DOI: https://doi.org/10.7208/chicago/9780226051826.001.0001

Ronan, P. (Ed.) (2016). *Perspectives on English in Switzerland.* Lausanne, Switzerland: University of Lausanne.

Sagan, C. (1987). Keynote address to the Committee for the Scientific Investigation of Claims of the Paranormal (today known as the Committee for Skeptical Inquiry).

Sutherland, S. (2013). *Irrationality: The enemy within.* London: Pinter & Martin Ltd.

Takakuwa, M. (2005). Lessons from a paradoxical hypothesis: A methodological critique of the threshold hypothesis. In J. Cohen, K. T. McAlister, K. Rolstad, & J. MacSwan (Eds.), *ISB4: Proceedings from the 4th International Symposium on Bilingualism* (pp. 2222–2232). Somerville, MA: Cascadilla.

Vanhove, J., & **Berthele, R.** (2015). The lifespan development of cognate guessing skills in an unknown related language. *International Review of Applied Linguistics in Language Teaching, 53*(1), 1–38. DOI: https://doi.org/10.1515/iral-2015-0001

Wattendorf, E. (2006). *Effects of early second language acquisition on the cortical language network in multilinguals. Evidence from fMRI.* Fribourg/Freiburg, Fribourg.

Willingham, D. T. (2012). *When can you trust the experts? How to tell good science from bad in education.* San Francisco, CA: Jossey-Bass.

Wulff, S., & **Ellis, N. C.** (2018). Usage-based approaches to second language acquisition. In D. Miller, F. Bayram, J. Rothman, & L. Serratrice (Eds.), *Bilingual cognition and language: The state of the science across its subfields* (pp. 37–56). Amsterdam: John Benjamins. DOI: https://doi.org/10.1075/sibil.54.03wul