View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by White Rose Research Online

This is a repository copy of *Who watches the watchmen? Evaluating evaluations of El Sistema*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/132193/

Version: Accepted Version

Article:
Baker, G., Bull, A. and Taylor, M. orcid.org/0000-0001-5943-9796 (2018) Who watches the watchmen? Evaluating evaluations of El Sistema. British Journal of Music Education, 35 (3). pp. 255-269. ISSN 0265-0517

https://doi.org/10.1017/S0265051718000086

This article has been published in a revised form in British Journal of Music Education [https://doi.org/10.1017/S0265051718000086]. This version is free to view and download for private research and study only. Not for re-distribution, re-sale or use in derivative works. © Cambridge University Press 2018.

Reuse
This article is distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs (CC BY-NC-ND) licence. This licence only allows you to download this work and share it with others as long as you credit the authors, but you can't change the article in any way or use it commercially. More information and the full terms of the licence here: https://creativecommons.org/licenses/

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Who watches the watchmen? Evaluating evaluations of El Sistema

Abstract

Within the growing field of publications on El Sistema and Sistema-inspired programmes around the world, a marked divide can be observed between the findings of critical academic studies and commissioned evaluations. Using evaluations of El Sistema in Venezuela and Aotearoa New Zealand as our principal case studies, we argue that this gulf can be explained at least partly by methodological problems in the way that some evaluations are carried out. We conclude that many Sistema evaluations display an alignment with advocacy rather than explorative research, and that the foundation for El Sistema's claims of social transformation is thus weak.

Keywords: **El Sistema; Sistema-inspired; evaluations; critical research; advocacy**

The Venezuelan National System of Youth and Children's Orchestras, better known as El Sistema, needs little introduction. Founded in 1975, it has become one of the largest and best-known music education programmes in the world. Over the last decade, its growing fame in the global North and the proliferation of 'Sistema-inspired' programmes in dozens of countries have led to an array of publications, which may be divided into three broad categories: (1) advocacy literature (e.g. Borzacchini, 2010; Tunstall, 2012; Tunstall and Booth, 2016); (2) critical academic studies (e.g. Logan, 2015a; Pedroza, 2015; Baker, 2016a, 2016b); and (3) commissioned evaluations (e.g. 'Evaluation of Big Noise,' 2011; Glasgow Centre for Population Health, 2015).

The critical academic studies have tended to focus primarily on ideological questions such as class, neoliberalism, and (neo)colonialism, scrutinising the programme from political, ethical, and historical perspectives (e.g. Borchert, 2012; Bull, 2016; Fink, 2016; Logan, 2016; Rosabal-Coto, 2016), though some studies have combined such critique with ethnographic research (e.g. Baker, 2014; Dobson, 2016). Commissioned evaluations, in contrast, have concentrated on more pragmatic and limited questions, which might be summarised as: does the programme work for current participants, and if so, in what ways? Surveying these two bodies of literature reveals a striking polarisation between the largely negative responses of critical scholars and the almost entirely positive conclusions of published evaluations.

This polarised scenario is the starting-point for our study. We ask: how might we account for the gulf between the findings of commissioned evaluations and critical writings (including published scholarship and blogs)?[1] We explore this question via two principal case studies, which involve detailed examination of evaluations of El Sistema in Venezuela and Aotearoa New Zealand, preceded by a brief historical introduction to the topic of evaluating this programme.

The history of evaluations of El Sistema in Venezuela

Divided opinions among researchers of El Sistema date back to the first attempts to evaluate the programme in 1996 to 1997, which were catalysed by El Sistema's efforts to secure funding from the Inter-American Development Bank (IDB). Four evaluations were produced by external consultants in this two-year period. The programme's efforts were successful: the IDB provided a Phase I loan of $8 million in 1998, and a Phase II loan of $150 million in 2008 – two of the most decisive developments in the history of El Sistema.

The first two reports, from 1996, were marked not only by a reverent tone but also by a striking lack of critical scrutiny or robust evidence of the supposed social benefits (see Baker with Frega, 2017). Rather than analysing the official narrative, they adopted it, emphasising the spiritual richness provided by music and its supposed capacity to overcome material poverty. Such was the advocacy tone that a segment of the report now forms part of El Sistema's official vision statement.[2] However, it appears that the IDB was not satisfied; it hired two more consultants and repeated the process the following year. The second pair of

consultants discovered numerous problems with the programme, which they documented in detail, revealing that the earlier evaluations had either missed or omitted many important issues and drawn dubious conclusions. Nevertheless, the IDB granted the $8 million loan, and the critical reports from 1997 were never made public.

The question marks that hovered over the robustness of the first two evaluations were not dispelled by subsequent studies. The next evaluation was carried out by the Universidad de los Andes in Mérida between 1999 and 2003. This quantitative study, too, reveals numerous flaws (Baker, 2014). As Hollinger notes (2006, 41–42), it has 'a number of inherent design weaknesses' and resembles 'less a scholarly endeavor than necessary documentation to advocate for The System.' As with the 1996 evaluations, the researchers adopted El Sistema's proselytising tone and in effect assumed an advocacy position.

A new evaluation by José Cuesta (2011) was used to justify the IDB's Phase II loan of $150 million. Yet it presented evidence of correlation rather than causation; the use of the terms 'treatment' and 'control' was misleading; it did not consider pre-existing cognitive or social differences between children; and El Sistema's leaders appeared to have played a part in creating the report (Baker, 2014). Furthermore, the financial calculations were questionable (Scruggs, 2015).[3] There are thus numerous reasons to doubt the study's speculative yet much-cited conclusion, a cost-benefit ratio of 1:1.68.

In fact, the IDB distanced itself from Cuesta's report. In 2011, the bank published a proposal for a new impact evaluation that would supposedly provide 'the first rigorous evidence of the results of the programme' ('Sistema Nacional,' 2011, 3). It also admitted that Cuesta's cost-benefit analysis 'was the result of various suppositions and not of a rigorous measurement of the impact of El Sistema' (2). (Nevertheless, it had already agreed the loan by this point.) This proposal gave rise to a large-scale experimental study, which was intended to finally settle the longstanding, unresolved question over the efficacy of El Sistema and put to rest the history of flawed evaluations, divided opinions, and changing views.

The 2016 IDB report: starting-points and conclusions

The new research was proposed in 2011, carried out in 2012-13, and first reported in 2016 (Alemán et al., 2016). The researchers created a 'theory of change' which hypothesised that 'short-term participation in orchestras or choruses may foster positive change in four child functioning domains: self-regulatory skills, behavior, prosocial skills and connections, and cognitive skills.' To test their theory, they measured 26 primary outcome variables within these 4 domains. Only two significant outcomes (at the 90% level) were found: 'the early-admission group had higher self-control and fewer behavioral difficulties, based on child reports.' There were thus no significant outcomes in 24 out of 26 areas, and the researchers 'did not find any full-sample effects on cognitive skills […] or on prosocial skills and connections.'

Perhaps more strikingly, the estimated poverty rate among the El Sistema participants was 16.7%, while the rate for the states in which they lived was 46.5%. In other words, the El Sistema participants in the experiment were three times less likely to be poor than all 6 to 14 year-olds residing in the same states. Consequently, the study 'highlights the challenges of targeting interventions towards vulnerable groups of children in the context of a voluntary social program.' Furthermore, 44% of students who were offered a place failed to complete two semesters. The study thus found little evidence to support the theory of change, but did find two statistics that raised doubts about the official narrative of El Sistema as a programme aimed primarily at, and with transformative effects on, the poor.

Looking at the genesis of this study is also revealing. A news article on the IDB's website, announcing the decision to undertake the research, opens with the words: 'When the first orchestra for young people from low-income families from the most deprived neighborhoods in Caracas was founded back in 1975.'[4] In fact, the social composition of the first orchestra was predominantly middle-class, and most participants were conservatoire students (Baker, 2014). Reproducing a myth is hardly a promising start. The article continues with another myth, describing the programme as 'giving priority to those from the lowest socio-economic levels'; yet El Sistema has no systematic targeting mechanisms, affirmative action policies, or quotas – hence the low poverty rate discovered by the IDB's subsequent research. These are surprisingly misinformed statements about fundamental aspects of the programme, considering that they come from its major non-state funder, and they suggest

that the IDB's starting-point was closer to the advocacy literature than to scholarly research.

The IDB's formal proposal document details the goals of the study: 'The expected impacts of El Sistema include the development of social skills and self-esteem, a reduction in the school dropout rate, particularly in secondary school, a reduction in the incidence of risky behaviours, and reduced frequency of unplanned pregnancies' ('Sistema Nacional,' 2011, 2). The second to fourth categories are emphasised repeatedly: on the following page, the objectives of the study are stated as 'to seek to generate rigorous evidence of the social effects of […] El Sistema, including the impacts on school dropout, illegal behaviour, and unplanned pregnancies' (3). Under 'expected results,' we read: 'The data will be used to evaluate rigorously the impacts of El Sistema on school dropout, risky behaviours, incidence of crime, and prevalence of unplanned pregnancies' (ibid.).

However, the study itself, published five years later, does not discuss rates of school dropout, crime, or unplanned pregnancies, nor does it mention this major shift in the evaluation's targets and aims. This omission raises significant questions. What happened to the issues identified as central in the proposal? When and why did they disappear?

The 2016 IDB report: methods

At this point, we move to a critical analysis of the 2016 IDB report itself. The study makes a number of claims about the effects of El Sistema using statistical

language; however, these claims fall far short of the standards that would ordinarily be used for the conclusions drawn. There are four key ways in which the report is deficient: preregistration, the use of the 90% significance level, p-hacking, and subgroup analysis. Each of these issues is sufficient to raise questions about the validity of the report; in combination, they undermine the analysis entirely.

The study was registered at clinicaltrials.gov. However, its registration falls short of best practice in two ways. The first is that the registration on clinicaltrials.gov was first received in 2015, and last verified in 2016. Given that the data was collected during 2012 and 2013, this is not preregistration, which is considered best practice in medical research (see e.g. de Angelis et al., 2004). The second issue is that the analysis is not specified: while outcome measures are listed, detail is not provided, and measures are specified in five groups rather than as a number of individual items, while the analytical technique is absent. Consequently, it is unclear whether the key issues in the proposal document were discarded before the data were collected or afterwards (perhaps because no significant effects were found). Also missing from the registration was any subgroup analysis: the only specified arms of the study were the treatment and control groups, with no suggestion that differences would be identified within any smaller groups (see below).

In the paper itself, instead of comparing five composite measures (as implied in the trial registration), the authors compare the treatment and control groups across 26 different outcome measures, using 90% as a threshold for statistical

significance. Opting for 90% is highly unusual. The overwhelming majority of papers that use significance testing employ the 95% significance level, and this itself has been frequently criticised as over-generous, with 99% or 99.9% being more appropriate for robust results. A single study with results significant at the 95% level would not ordinarily be sufficient to justify a policy intervention or a change in prescribing policy; it is highly unusual that a large organisation such as the IDB would use the weaker 90% threshold without a clear justification.

Differences that are addressed in the conclusion are first seen in the 'Impacts' section. The authors find differences between the treatment and control group in 2 out of 26 measures. If the authors were using no adjustment, given the large number of outcomes measured, and there were no genuine underlying differences, the probability of at least one difference significant at the 95% level is 74%: while this may seem unusually high, it can be understood by comparing the cumulative probability of each and every difference being nonsignificant. The probability of at least one difference significant at the 90% level is 92%. Therefore, the authors' discovery of differences between young people participating in El Sistema is almost trivial.

However, the authors acknowledge the large amount of hypothesis-testing, which would ordinarily yield significant differences through sheer luck, and report that they control the k-familywise error rate – that is, the probability of observing a given number (k) of false positives, in the event that any true differences observed in the data were due to random noise – with an adjusted version of the Romano-Wolf procedure, a technique to adjust thresholds for

significance testing given the number of comparisons being drawn simultaneously. The adjustment is that instead of using k =1, which they find too conservative, they use k=h/2 (where h = the number of variables in a domain). While the authors they cite (Delattre and Roquain, 2015) do demonstrate that the Romano-Wolf procedure is conservative, they do not demonstrate that k=h/2 is an appropriate adjustment. In addition, the measures in which they investigate differences are likely to be correlated, so while the authors are to be commended on adjusting their tests to acknowledge the numbers of tests they have conducted, their adjustment is far too generous, implying levels of significance unlikely to be upheld by their results. That the results they do yield are only significant at the 90% level (and only 2 out of 26 tests at that) does not provide support for the alternative hypothesis that there are differences between the students participating in El Sistema for a year and those who waited a year.

These differences between those participating in El Sistema and the apparent control group are not the only differences addressed in the conclusion. The authors also compare these groups *within subgroups*. These are broken down as follows. First, they compare whether students' mothers have any college education or not; second, they compare students aged between 6 and 9, and between 10 and 14; third, they compare the interaction between gender and exposure to violence (that is, comparing boys who have been exposed to violence with those who have not, with girls who have, and with girls who have not). With 8 different groups compared across 26 different measures, this leads to a total of 208 comparisons between treatment and control groups. The paper only reports coefficients that are significant at the 90% level, of which there are 13.

According to the experimental literature, subgroup analysis 'can lead to overstated and misleading results' (Wang et al., 2007, 2189), and this is particularly the case in the event of multiple subgroups being analysed, a problem known as multiplicity. Here, there are eight different subgroups being analysed. While the experimental literature strongly advises that any subgroup analysis be registered before data is collected and analysed, this is not sufficient to solve the multiplicity problem where – again – significant results are likely to appear through sheer luck. In this case, there is no reference to subgroup analysis in the trial registration. A critical reader might wonder how many other subgroups have been analysed and discarded before these eight were settled on. Why analyse *together* boys and girls whose mothers are more educated, but *separately* boys and girls who have been exposed to violence? At least, as the authors are really testing an interaction effect here, any differences identified should be net of the direct effect: instead of comparing boys and girls who have been exposed to violence, it should be clear what the effect of exposure to violence on each outcome measure is. However, the direct effect of exposure to violence is not reported. Given the absence of preregistration, it seems likely that all possible subgroup analyses were conducted; it is not computationally intensive and there is no clear reason for choosing these subgroups rather than others. This is a technique known as p-hacking (see e.g. Head et al., 2015; Bruns and Ioannidis, 2016). It is therefore probable that the use of the k-familywise error is insufficiently rigorous, and the apparent significant differences represent random noise. Furthermore, it appears likely that this study incorporated a 'fishing exercise,' investigating thousands of dimensions in which

differences between participants and non-participants in El Sistema might exist, and making no mention of having investigated the overwhelming majority of those dimensions where El Sistema was shown not to have made any difference.

These issues, taken together, make the report almost impossible to take seriously. Three of its authors are employees of the IDB, which had been funding El Sistema for the previous 18 years and cannot therefore be considered an impartial observer. The trial was not preregistered; the threshold to which the authors ascribe significance is half as demanding as the academic mainstream, yet one that they mostly fail to reach; and the number of analyses run implies that apparently significant results are likely to be the result of dumb luck. At the very least, the study's conclusions that 'exposure to El Sistema might serve an important role as a preventive strategy to promote positive outcomes among disadvantaged children' and 'El Sistema is particularly effective for vulnerable males' are notably overstated. But we suggest that the report in fact represents a form of cargo cult analysis: it is full of superficially technical and analytical work, but it needs only the gentlest of interrogation to reveal that is built on sand.

Finally, there is a potential generalisation problem given the application process. Even if one were to accept that El Sistema had been proven to have beneficial effects on participants, the conclusions do not take account of the fact that all participants were signed up for the programme by a parent or guardian, who thus showed a certain level of commitment to the child's education. It cannot be assumed that effects on children from more supportive families will be mirrored in all children; there is no evidence that the study has external validity. It may be

that parental support is a key ingredient in generating positive effects – something suggested by studies of El Sistema and other after-school programmes (Pérez and Rojas 2013; Baker 2014; Cid 2014). This would limit the wider applicability of El Sistema as a social inclusion programme, and if we also take into account the IDB's findings about poverty and dropout rates, El Sistema may in fact be quite ineffective in promoting positive change in the most disadvantaged sectors of society.

Evaluating Sistema Aotearoa

Evaluations of non-Venezuelan Sistema programmes have generally been assumed to be robust, and having been cited frequently in the media, they play an important role in advocacy for the global Sistema movement. However, Owen Logan's (2015b) scathing assessment of two evaluations of Sistema Scotland, followed up by Baker (2017), suggests that more critical scrutiny would be worthwhile. An evaluation of Sistema Aotearoa, a government-funded programme that began in April 2011 in Auckland, New Zealand, is the focus of our second case study. The Ministry for Culture and Heritage and Auckland Philharmonia Orchestra, who jointly run Sistema Aotearoa, commissioned the Kinnect Group, a private sector evaluation consultancy based in Aotearoa New Zealand,[7] to evaluate this programme. They produced an initial report in 2012 and an 'outcome evaluation' three years later (McKegg et al., 2015).

The evaluation draws on two sources: quantitative educational achievement data and qualitative 'success case studies' from participants and their families. The

quantitative aspect of the evaluation draws on data from 'overall teacher judgements' (OTJs), which are part of the national standards for statutory education that were introduced in 2012 (Thrupp, 2013). This policy has provoked numerous concerns, for example that using teachers' judgements of pupils may exacerbate existing inequalities (see Thrupp and Easter, 2012; Thrupp and White, 2013). Additionally, the report is frank about further limitations of this data. These include the lack of baseline data from when the programme began; lack of data for three out of the seven Sistema Aotearoa schools; lack of data about children who dropped out of the programme; and use of aggregate data at the level of the school rather than individual pupils' data. These issues, as well as the small size of the data sets, mean that any conclusions drawn from this data are partial and very limited. For the two years of data that are available, the study finds a statistically significant improvement in reading and maths achievement, but notes that 'it is possible that the difference we have identified is because the higher achieving students are more likely to stay engaged with the programme' (McKegg et al., 2015, 17). Given a dropout rate of 47%, this is an important caveat.

Since only very tentative conclusions can be drawn from the quantitative data, the qualitative element of the study is very important. However, the 'success case' methodology that is used is problematic. This approach is not a well-known social science methodology, and is not mentioned in the most reputable book on case study methods, by Robert K. Yin (2013). Nor does it appear in other relevant textbooks (e.g. Gerring, 2010; Woodside et al., 2006). Stufflebeam and Coryn (2014) note that it has certain strengths, such as identifying what works

well, reassuring funders, and boosting morale, and it is also quick and cheap. However, it has obvious weaknesses: it is narrow, short-term in outlook, suffers seriously from selection bias, and does not present 'a comprehensive assessment of a programme's merit and worth' (ibid., 143) (see Baker 2016c).

For the Sistema Aotearoa evaluation, this method involved recruiting five pupils who were identified by teachers as being particularly successful in the programme and carrying out interviews with them and their families, as well as with Sistema staff. The reason given for adopting the 'success case' methodology is cultural sensitivity. The programme involves a high number of Pacific Island families, and the report argues that 'it is considered impolite in Pacific cultures to talk in negative ways about a service or programme that is being received,' which therefore suggests that a 'success case' approach will 'generate responses that are both culturally valid and more accurate' (McKegg et al., 2015, 9-10). This is indeed a methodological hurdle, but one that is insufficient to justify such a partial approach. More imaginative methodological choices, such as Rimmer, Street, and Phillips's (2014) creative methods with children, could have overcome this issue.

This 'success case methodology' means that evidence within the data of less positive outcomes of the programme is not discussed. These include the dropout rate of 47%; the major intervention in family life that Sistema Aotearoa requires; the experience of stigmatisation described by some parents in the programme when attending prestigious concert venues; and signs that the programme adopts a deficit model of culture. To briefly discuss the last of these, the report

notes that 'orchestral music typically falls outside of what these children […] see as "their culture"' but also that through participation in Sistema Aotearoa, children and their families 'are all gaining an appreciation for Western orchestral musical culture that they otherwise would not have had the opportunity to do' (ibid., 31-32). However, the report recommends that the programme outcomes acknowledge that Māori and Pacific Islanders are not simply recipients of knowledge from the programme, but already have existing knowledge and practices. In fact, public and media discussions in New Zealand have been more critical than the evaluation, showing concern that Sistema Aotearoa may be devaluing Māori and Pacific Island culture (McPhail et al., 2018, 4; Trinik, 2014, 14). These discussions reflect an awareness that Aotearoa New Zealand is still deeply shaped by its history of colonisation. Stark divides continue to exist in health, education, and social outcomes between Pākehā (white New Zealanders) and Māori and Pacific Islanders (Statistics New Zealand, 2016). Therefore a cultural education programme in which colonised people learn to 'gain an appreciation' for European culture requires contextualisation within a wider critical discussion, which the 'success case' methodology adopted in this evaluation does not allow.

In sum, while all social data is necessarily partial, the data in this study is particularly limited. Most problematic is the presentation of the evaluation as an 'outcome evaluation' when it should more accurately be described as a collection of accounts from a handful of participants who enjoyed the programme. Missing are critical discussions of negative and null outcomes that inevitably occur as well, and of the wider context of class and race inequality in which Sistema

Aotearoa operates, in particular how an education programme that teaches European high culture to Māori and Pacific Islanders should be understood in the context of Aotearoa New Zealand's colonial legacy. This is where academic research diverges from commissioned evaluations. Indeed, Aotearoa New Zealand academics lead the world in 'decolonising methodologies', examining ways of re-thinking how to do research with colonised groups (Tuhiwai Smith, 1999). Proper assessment of a programme that includes in its aims an explicit transfer of the knowledge and culture of white Europeans towards groups who have been colonised by them necessitates a critical examination of this context.

The uses and abuses of Sistema evaluations

The flaws and limitations that have been found in evaluations of El Sistema, both in Venezuela and elsewhere, give cause for concern. They raise doubts about both the efficacy of El Sistema and the processes of evaluating such programmes. These concerns and doubts are only amplified by considering the post-publication trajectory of two of these studies.

Cuesta's (2011) principal conclusion, a cost-benefit ratio of 1:1.68, quickly became a mainstay of advocacy arguments in favour of El Sistema and Sistema-inspired programmes. It exemplifies Belfiore's (2016, 212) statement, drawing on Max Singer's article 'The vitality of mythical numbers,' that 'once a statistic is produced (no matter whether rigorously or incorrectly) and starts being quoted, it takes on a life of its own. As a result, the imaginary statistics might enter the official debate on cultural policy, being quoted for years without its original

source and its reliability ever being verified.' Four years later, the IDB very quietly recognised the speculative nature of Cuesta's conclusion, and more thorough academic critiques followed.[8] However, by this point the figure's work was already done: it had underpinned both the IDB's decision to issue a $150-million loan and advocacy arguments for establishing Sistema-inspired programmes around the world. The story of Cuesta's report illustrates that headline numbers may garner far more attention than the detail of the studies that support them, even with such large sums at stake, and that questions may come too late to make any difference and/or be ignored by interested parties (for example, Tunstall and Booth (2016, 228) employ this figure despite knowing that it had been criticised).

Even more striking has been the official dissemination of the 2016 report. Although the conclusions of this study were overstated in relation to the findings, the authors did signal two important negative findings – the low poverty rate and high dropout rate – and were open about having found no significant outcomes in 24 out of 26 areas, concluding: 'We did not find any full-sample effects on cognitive skills […] or on prosocial skills and connections.'

The IDB's blog post on the report, though, gave it a much more positive spin.[9] It mentioned none of the negative or equivocal findings, only the positives. It even made positive claims about gender equality, whereas the study itself had found the opposite. The misleading impression given by the blog post was that the study provided an unequivocal stamp of approval for El Sistema.

A launch event for the study in Caracas in March 2017, at which the report's authors, El Sistema leaders, and government representatives were present, continued in this vein. The press release declared that the research team 'expressed its satisfaction with the possibility of confirming the transformative work of the programme.'[10] The researchers had concluded, it claimed, that the children and young people in El Sistema showed improved connections with school and family, a higher degree of cooperation with their peers, and greater self-confidence. According to one, Marco Stampire, 'we found a decrease in levels of aggression and risk-taking […]; and a willingness to take part in collective activities. The positive effects were also manifested in childhood IQ.' These claims contradicted the evidence and conclusions presented by the same researchers in their published article, in which they had stated that they had not found any full-sample effects on cognitive skills or prosocial skills and connections.

Ferdinando Regalía, head of the IDB's Social Protection and Health Division, stated that the results 'tackle the criticisms of El Sistema's work and reaffirm the value of social inclusion via a programme of artistic and musical education.' Yet the findings about the poverty and dropout rates did not tackle such criticisms but rather confirmed their validity.

On the basis of the press release, it is impossible to be sure whether the IDB research team overstated the positives and omitted the negatives from its public presentation, or whether El Sistema's press office was responsible. Either way, there are reasons to be concerned about the way this study is being used. The

press release suggested that the official line from the institutions involved was that the study proved El Sistema to be a success, and this impression was confirmed by subsequent publications such as a graphic summarising the study, released by the IDB, and an interview with El Sistema's executive director, Eduardo Méndez, both of which gave the findings an entirely favourable spin.[11] If the positive conclusions of the report were overstated in the first place, they were subsequently exaggerated further in its public presentation and shorn of important caveats.

The institutions involved all have good reason to portray the findings in the best possible light. The IDB and the Venezuelan government have invested hundreds of millions of dollars in El Sistema over a period of many years, and the unvarnished findings of the report provide little justification for this expenditure. The pressure on the researchers must therefore have been intense, which might explain the generous statistical approach used. But the public presentation of the report raises serious questions about the value of investing significant resources over a period of years in evaluating a large programme with major, longstanding support from politicians and multinational institutions. Similarly, the Sistema Aotearoa evaluation is designed in such away as to *avoid* bringing to light any problems or criticisms associated with the programme. Furthermore, in a country where recognition of Māori culture is enshrined in law (New Zealand Law Commission, 2001), it is highly unusual that the evaluation failed to discuss the effects of the programme on the cultural values of its predominantly Māori and Pacific Island participants. Both the Sistema Aotearoa and the IDB reports, therefore, and above all the associated publicity for the

latter, appear more like rubber-stamping exercises or even a whitewash than a serious attempt to identify strengths and weaknesses and to improve the programmes accordingly.

Conclusion

We began our study with the problem of the marked divide within the Sistema research literature. There are a number of possible explanations for this scenario, including the disciplinary backgrounds and 'homes' of the researchers, and the organisations for which they work (for example, universities or consultancies). Also, one obvious reason that the two sub-fields produce quite different answers is that they ask quite different questions. Whereas evaluators tend to examine whether programmes achieve their goals, independent researchers are much more likely to interrogate the validity of those goals and consider cultural, political, or philosophical questions that they raise, drawing on academic fields such as music studies, sociology, and critical theory.[12]

While this distinction is important, and indeed worthy of a separate study, here we have sought to shed light on our central question by shifting the focus of critical enquiry away from El Sistema and its spinoffs and towards the evaluations of these programmes. We conclude that the gulf within the literature may relate, at least in part, to flaws in the processes of some evaluations, which lead them to present an overly optimistic picture. Through our critique of these two studies, we suggest that the research foundation for El Sistema's claims of social transformation, and hence for its fame and international proliferation, may

be weaker than first appears and therefore require further examination. (Indeed, a striking aspect of El Sistema's history is that repeated investigations in Venezuela over a period of twenty years have failed to generate robust evidence of its efficacy.) We are not making an a priori claim that all evaluations are flawed or that independent academic research is necessarily superior, but rather suggesting that evaluations (like all research) deserve careful scrutiny, and that when looking for explanations for gaps between the conclusions of evaluations and academic studies, one route to explore is critical revision of evaluative methodologies.

As a further step, it is illuminating to consider Eleonora Belfiore's call for a 'critical research ethos,' which she defines as

> research that is disinterested, that is, indifferent to the requirements of advocacy – advocacy being a fully legitimate enterprise, but one completely distinct and, ideally, separate from genuinely explorative research. By 'explorative' research, I refer to a type of research that aims to describe, explore and illuminate complex issues around the role and condition of culture, cultural production, consumption and administration in contemporary society. (Belfiore, 2009, 354)

We suggest that the flaws in some Sistema evaluations are linked to the fact that many such studies – the 1997 reports on the Venezuelan programme being an important exception – display an alignment with advocacy rather than explorative research, and are examples of what Logan (2015b) calls "Sistema-

friendly research." This alignment takes a variety of forms, which include overplaying positive findings, underplaying negative ones, side-lining problematic issues, omitting reference to the critical research literature on El Sistema specifically and music education more generally, and/or adopting rather than scrutinising the programme's official rhetoric and proselytising tone. These features are characteristic of the advocacy literature referenced at the start of the article. There is thus a reproduction *within* the research field of the division between independent academic studies and non-academic advocacy writing.

Our primary aim has been to shed critical light on the role of evaluations in reinforcing, rather than genuinely testing, the excessively optimistic dominant narrative about El Sistema. However, we also hope that our research will make a contribution to debates within the field of cultural policy studies, which has been raising questions about evaluations of the social impact of the arts for a number of years (e.g. Merli, 2002; Belfiore, 2002; Selwood, 2003; Belfiore, 2009; Belfiore and Bennett, 2010; Lees and Melhuish, 2015; Johanson and Glow, 2015). Finally, we also suggest that our study should encourage critical debate on the topic of evaluations in music education, such as reflection on what constitutes robust forms of evidence, the relationship between evaluation, advocacy, and explorative research, and ways in which evaluation data can be misrepresented and misreported. We hope to have shown that programme evaluations are a valid and worthwhile object of critical research. Yet, ideally, this debate should include but also go beyond the sort of scrutiny of existing evaluations that we have undertaken here and penetrate the commissioning and evaluation

processes themselves, since post-hoc scrutiny is unlikely to affect policy and funding decisions.

We would also suggest that evaluations should be seen as political tools, in the sense that they frequently operate within challenging funding contexts in which there are losers as well as winners. As Bull (2016, 140) notes, In Harmony El Sistema England received considerable investment at a time when music education funding generally was being cut by nearly a third in England, and Sistema Scotland has flourished against a similarly concerning backdrop (Baker 2017). Excessively optimistic El Sistema evaluations such as those studied in this article may therefore have implications for the wider field of music education, potentially diverting resources and/or attention away from more effective or equitable programmes.

If critical findings are made clearly visible in reports' conclusions and are acknowledged by programmes, then evaluations may be a valuable spur to positive change in music education. But if such findings are played down in reports and then airbrushed out of the picture in their public presentation – as occurred with the 2016 IDB study, but can also be seen in evaluations, publicity, and media stories about Sistema Scotland – then evaluations simply serve as a justification for the status quo, however problematic it may be.

6050 words (without references)

Bibliography

ALEMAN, X. et al. (2016) The Effects of Musical Training on Child Development: A Randomized Trial of *El Sistema* in Venezuela. *Prevention Science*.

BAKER, G. (2014) *El Sistema: Orchestrating Venezuela's Youth*. New York: Oxford University Press.

BAKER, G. (2016a) Editorial Introduction: El Sistema in critical perspective. *Action, Criticism & Theory for Music Education*, **15 (1)**, 10–32.

BAKER, G. (2016b) Citizens or Subjects? El Sistema in Critical Perspective. In D. Elliott, W. Bowman, & M. Silverman (Eds), *Artistic Citizenship: Artistry, Social Responsibility, and Ethical Praxis* (pp. 313-38). New York: Oxford University Press.

BAKER, G. (2016c) Antes de pasar página: conectando los mundos paralelos de El Sistema y la investigación crítica. *Revista Internacional de Educación Musical*, **4**, 51-60.

BAKER, G. (2017) Big noise in Raploch? *Scottish Review*, 21 June. http://www.scottishreview.net/GeoffBaker285a.html.

BAKER, G. & FREGA, A. L. (2017) Los reportes del BID sobre El Sistema: Nuevas perspectivas sobre la historia y la historiografía del Sistema Nacional de Orquestas Juveniles e Infantiles de Venezuela. *Epistemus*.

BELFIORE, E. (2002) Art as a Means of Alleviating Social Exclusion: Does It Really Work? A Critique of Instrumental Cultural Policies and Social Impact Studies in the UK. *International Journal of Cultural Policy*, **8 (1)**, 91–106.

BELFIORE, E. (2009) On bullshit in cultural policy practice and research: notes from the British case. *International Journal of Cultural Policy*, **15 (3)**, 343-359.

BELFIORE, E. & BENNETT, O. (2010) Beyond the 'Toolkit Approach': Arts Impact Evaluation Research and the Realities of Cultural Policy-Making. *Journal for Cultural Research*, **14 (2)**, 121-142.

BELFIORE, E. (2016) Cultural policy research in the real world: Curating 'impact', facilitating 'enlightenment'. *Cultural Trends*, **25 (3)**, 205-216.

BORCHERT, G. (2012). Sistema Scotland: A Critical Inquiry into the Implementation of the El Sistema Model in Raploch. MMus, University of Glasgow.

BORZACCHINI, C. (2010) *Venezuela en el cielo de los escenarios*. Caracas: Fundación Bancaribe.

BRUNS, S. B. & IOANNIDIS, J. P. A. (2016) *p*-Curve and *p*-Hacking in Observational Research. *PLoS ONE*, **11 (2)**, e0149144.

BULL, A. (2016) El Sistema as a Bourgeois Social Project: Class, Gender, and Victorian values. *Action, Criticism, and Theory for Music Education*, **15 (1)**, 120–53.

CID, A. (2014) Giving a second chance: an after-school programme in a shanty town interacted with parent type: lessons from a randomized trial. *Educational Research and Evaluation*, **20 (5)**, 348-365.

CUESTA, J. (2011) Music to My Ears: The (Many) Socioeconomic Benefits of Music Training Programs. *Applied Economics Letter*, **18 (10)**, 915-18.

DE ANGELIS, C. et al. (2004) Clinical Trial Registration: A Statement from the International Committee of Medical Journal Editors. *The New England Journal of Medicine*, **351**, 1250-1251.

DELATTRE, S. & ROQUAIN, E. (2015) New procedures controlling the false discovery proportion via Romano-Wolf's heuristic. *The Annals of Statistics,* **43**, 1141–1177.

DOBSON, N. (2016) Hatching Plans: Pedagogy and Discourse Within an El Sistema-Inspired Program. *Action, Criticism, and Theory for Music Education*, **15 (1)**, 89–119.

ELPUS, K. (2015) Music Teacher Licensure Candidates in the United States: A Demographic Profile and Analysis of Licensure Examination Scores. *Journal of Research in Music Education*, **63 (3)**, 314-335.

EVALUATION OF BIG NOISE, SISTEMA SCOTLAND. (2011). Scottish Government Social Research.

FINK, R. (2016) Resurrection Symphony: El Sistema as Ideology in Venezuela and Los Angeles. *Action, Criticism, and Theory for Music Education*, **15 (1)**, 33–57.

GERRING, J. (2006) *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.

GLASGOW CENTRE FOR POPULATION HEALTH. (2015). Evaluating Sistema Scotland – initial findings report. Glasgow: GCPH.

HEAD, M. L., HOLMAN, L., LANFEAR, R. & JENNIONS, M. D. (2015) The Extent and Consequences of P-Hacking in Science. *PLoS Biol*, **13 (3)**, e1002106.

HOLLINGER, D. (2006) Instrument of Social Reform: A Case Study of the Venezuelan System of Youth Orchestras. DMA diss., Arizona State University.

HOLOCHWOST, S., WOLF, D. P. & BOSE, J. H. (2017) Building Strengths, Buffering Risk: Evaluating the Effects of El Sistema-Inspired Music Programs in the United States. http://wolfbrown.com/images/books_reports/Building_Strengths_Buffering_Risk.pdf.

JOHANSON, K. & GLOW, H. (2015) A virtuous circle: the positive evaluation phenomenon in arts audience research. *Participations*, **12 (1)**, 254–270.

LEES, L. & MELHUISH, C. (2015) Arts-led regeneration in the UK: The rhetoric and the evidence on urban social inclusion. *European Urban and Regional Studies*, **22(3)**, 242–260.

LOGAN, O. (2015a) Doing Well in the Eyes of Capital: Cultural Transformation from Venezuela to Scotland. In J.-A. McNeish, A. Borchgrevink and O. Logan (Eds), *Contested Powers: The Politics of Energy and Development* (pp. 216–253). London: Zed Books.

LOGAN, O. (2015b) Hand in glove: El Sistema and neoliberal research. https://www.researchgate.net/publication/287202150.

LOGAN, O. (2016) Lifting the Veil: A Realist Critique of Sistema's Upwardly Mobile Path. *Action, Criticism, and Theory for Music Education*, **15 (1)**, 58–88.

MCKEGG, K., CROCKET, A., GOODWIN, D., & SAUNI, P. (2015) Sistema Aotearoa: Outcomes evaluation report. Hamilton: The Knowledge Institute Limited (Kinnect Group).

MERLI, P. (2002) Evaluating the social impact of participation in arts activities: A critical review of François Matarasso's *Use or Ornament? International Journal of Cultural Policy*, **8 (1)**, 107-118.

NEW ZEALAND LAW COMMISSION. (2001) *Māori Custom and Values in New Zealand Law*, Study Paper 9, Wellington, N.Z: The Law Commission.

PEDROZA, L. (2015) Of Orchestras, Mythos, and the Idealization of Symphonic Practice: The *Orquesta Sinfónica de Venezuela* in the (Collateral) History of El Sistema. *Latin American Music Review*, **36 (1)**, 68-93.

PEREZ, E. & ROJAS, Y. (2013) ¿Por qué quiero que mi hijo sea músico? Expectativas de las madres, cuyos hijos están en la OSIC. Diss, Universidad Católica Andrés Bello. http://biblioteca2.ucab.edu.ve/anexos/biblioteca/marc/texto/AAS7348.pdf

RIMMER, M., STREET, J. & PHILLIPS, T. (2014) Understanding the cultural value of In Harmony-Sistema England. Arts & Humanities Research Council/University of East Anglia.

ROSABAL-COTO, G. (2016) Costa Rica's SINEM: A Perspective from Postcolonial Institutional Ethnography. *Action, Criticism, and Theory for Music Education*, **15 (1)**, 154–87.

SCRUGGS, T. M. (2015) 'The Sistema,' the Euroclassical Tradition, and Education as a Transformative Agent to Supercede Class Status. Paper presented at the annual meeting of the Society for Ethnomusicology, Austin, Texas, December 3-6.

SELWOOD, S. (2002) Measuring culture. *Spiked Online*. http://www.spiked-online.com/newsite/article/6851#.WOd4UGTyugQ.

SISTEMA NACIONAL DE ORQUESTAS JUVENILES E INFANTILES. (2011) Evaluación de Impactos. http://idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=36583351.

STATISTICS NEW ZEALAND. (2016). Life expectancy. http://www.stats.govt.nz/browse_for_stats/snapshots-of-nz/nz-social-indicators/Home/Health/life-expectancy.aspx (accessed 4.28.17).

STUFFLEBEAM, D. L. & CORYN, C. L. S. (2014) *Evaluation Theory, Models, and Applications* (2nd edition). San Francisco: Jossey-Bass.

THRUPP, M. (2014) At the eye of the storm: Researching schools and their communities enacting National Standards. *New Zealand Journal of Educational Studies*, **49 (1)**, 6-20.

THRUPP, M. & EASTER, A. (2012) Research, analysis and insight into National Standards (RAINS) Project: First Report Researching Schools' Enactments of New Zealand's National Standards Policy. http://www.nzei.org.nz/site/nzeite/files/reports/RAINS-Final-2012-03-01.pdf

THRUPP, M. & WHITE, M. (2013) Research, analysis and insight into National Standards (RAINS) Project. Final report: National Standards and the Damage Done. Wellington, New Zealand: The New Zealand Educational Institute Te Riu Roa (NZEI). http://www.education2014.org.nz/

TRINIK, R. (2014) Window with a view: reflections on Sistema Aotearoa. *E-journal of studies in music education*, **10 (1)**, 12–21. http://www.arts.canterbury.ac.nz/music/merc/downloads/ejournal_v10no1.pdf

TUHIWAI SMITH, L. (1999) *Decolonizing Methodologies: Research and Indigenous Peoples*. London: Zed Books.

TUNSTALL, T. (2012) *Changing Lives: Gustavo Dudamel, El Sistema, and the Transformative Power of Music*. New York: W. W. Norton.

TUNSTALL, T. & BOOTH, E. (2016) *Playing for Their Lives: The Global El Sistema Movement for Social Change Through Music*. New York: W. W. Norton.

WANG, R. et al. (2007) Statistics in Medicine — Reporting of Subgroup Analyses in Clinical Trials. *The New England Journal of Medicine*, **357**, 2189-2194.

WOODSIDE, A. (2010) Case Study Research: Theory, Methods and Practice. Bingley: Emerald Group Publishing.

YIN, R. Y. (2013) *Case Study Research: Design and Methods* (5th edition). Los Angeles: SAGE Publications.

---

[1] Key critical blogs include http://laotracaradelsistema.blogspot.com.co/; https://jonathangovias.com; and https://geoffbakermusic.wordpress.com/.

[2] http://fundamusical.org.ve/category/el-sistema/mision-y-vision/.

[3] For example, each robbery avoided, allegedly as a result of El Sistema's presence, is calculated as producing an economic benefit of $5,000 USD. As Scruggs notes, this figure is "rather fantastical" considering the economic realities of the neighbourhoods in which many beneficiaries reside.

[4] http://www.iadb.org/en/topics/social-protection/music-for-a-better-future,6964.html.

[7] As it is a bicultural country, both the Māori and English names for New Zealand are used.

[8] Baker (2014) and Scruggs (2015) were preceded in 2012 by https://geoffbakermusic.wordpress.com/el-sistema-older-posts/scam-voodoo-or-the-future-of-music-the-el-sistema-debate-2/.

[9] https://blogs.iadb.org/desarrollo-infantil/2016/12/15/musica/.

[10] http://fundamusical.org.ve/prensa/noticias/el-bid-confirma-impacto-positivo-de-el-sistema-en-ninos-y-jovenes/.

[11] http://blogs.iadb.org/wp-content/blogs.dir/35/files/2016/10/IE011-VE-MusicForDevelopment-ENG-2048.png; https://www.venezuelasinfonica.com/sistema-celebra-43-anos-atendiendo-mas-900-mil-ninos-jovenes-toda-venezuela.

[12] To take one example, a 2017 evaluation of Sistema-inspired programs in the United States recorded that nearly two-thirds of the students in the study were African American or Hispanic (Holochwost, Wolf, and Bose 2017), yet there was no discussion of the issue of race, even though 86% of music teacher licensure candidates in the US are white (Elpus 2015). This is precisely the kind of issue that has caught the attention of independent researchers working on El Sistema and Sistema-inspired programmes. On race-related criticisms raised by teachers at a prominent Sistema-inspired programme, see https://geoffbakermusic.wordpress.com/el-sistema-older-posts/playing-for-their-lives-sins-of-mission-and-omission-2/.