Replication in Music Psychology

Klaus Frieler,[1] Timo Fischinger,[2] Daniel Müllensiefen,[3] Kathrin Schlemmer,[4] Kelly Jakubowski,[3] and Kai Lothwesen[5]

[1] Department of Musicology, Hochschule für Musik Franz Liszt Weimar, Germany

[2] Institute of Music, Universität Kassel, Germany

[3] Department of Psychology, Goldsmiths College, London, UK

[4] Katholische Universität Eichstätt-Ingolstadt, Germany

[5] Hochschule für Musik, Frankfurt/M., Germany

Corresponding Author: Klaus Frieler, Department of Musicology, Hochschule für Musik Franz Liszt Weimar, Germany, klaus.frieler@hfm-weimar.de

]

Abstract

In this paper we address the current state and general role of replication in empirical sciences in general and music psychology in particular. We argue that replication should be an integral part of the quality management of science because it helps to improve and maintain the general quality of empirical sciences by enhancing the confidence in scientific phenomena and theories. Replicating empirical experiments has two major benefits: (1) It increases the sheer number of observations and (2) it provides independent evidence which works as a safety net against methodological fallacies, causally influential but unknown (i.e. random) factors, researcher degrees of freedom, and outright fraud. Furthermore, we argue that for low-gain/low-cost sciences such as music psychology, measures to ensure quality standards, in particular the amount of replication experiments conducted, can be expected to be lower than in high gain / high cost sciences. These lower expectations stem from the general acknowledgments that in low-gain/low-cost sciences (1) research resources are normally scarce and (2) the consequences of inadequate theories are relatively harmless. We argue that the view of music psychology as a low-cost/low-gain science can explain the striking lack of replication studies and meta-analyses. We also discuss possible counter-measures to enhance the reliability of music-psychological knowledge.

Replication in Music Psychology:

Science as a "belief" system

Replication is one of the core concepts of the scientific method because it increases the reliability of experimental results, thereby strengthening the confidence in scientific knowledge and, consequently, in scientific models and theories in general (for a general overview of different scientific paradigms as relevant to psychology see Dienes, 2008).

Confidence in a theory or, more generally, in any scientific proposition is enhanced and facilitated by certain factors (in no particular order): (1) *Expert knowledge*. If an acclaimed expert in the subject field advocates a theory, one is more likely to believe it. (2) *Peer pressure*. If one's peers subscribe to an opinion, one is more likely to believe it. (3) *Resonance/Plausibility*. If the stated theory resonates with one's own experience, introspective knowledge, ideology or one's other beliefs, one is more likely to believe it. (4) *Independent evidence*. If the same opinion is communicated from several (seemingly) independent sources, one is more likely to believe it. (5) *Empiricism.* The more a theory is based on empirically gathered data, the more one is likely to believe it. It may be argued that the first three points hold quite generally for any cultural and historical context, while the latter two might already be part of the program of empirical sciences devised within a specific (Western) context. Many examples can be recalled for which strong beliefs are not based on any independent empirical data (e.g., most religious beliefs). However, if one believes in the empirical program, experiments serve as an argument in a discourse with a high impact on belief structures.

Replication as part of scientific quality management

Replication of scientific experiments is an example of 'independent evidence'. There are two main ways through which replication can provide independent evidence: One is of a statistical nature, the other psychological, and both are interconnected. On the statistical side, it is well-established that estimates of parameters and probabilities become more precise (i.e., smaller standard errors) and are more reliable—hence more believable—with a growing number of observations. One consequence that may result from having a higher number of observations or replications is the so-called decline effect (Schooler, 2011), which occurs when a replication of an initially successful published experiment is less likely to yield an equally strong effect in subsequent replications. This may be because the initial finding overestimates the true effect size due to publication bias and the file drawer effect (Ferguson & Heene, 2012). The decline effect can mostly be accounted for by a regression to the mean (Galton, 1886; Nesselroade et al., 1980; Kahneman, 2011). On the psychological side, replication is one method of error detection and fraud prevention. After all, experiments are conducted by humans who have a variety of skills and motives, and who occasionally make mistakes or even intentionally manipulate data (Bakker & Wicherts, 2011, Stroebe et al., 2012; Simmons et al., 2011; Fanelli & Tregenza, 2009). Kahneman (2011) and others (e.g., Mahoney, 1977) have demonstrated that scientists are prone to cognitive biases as well. More gravely, the current incentive scheme of the scientific publication system, which rewards those presenting the most fresh, original, and surprising results, might seduce a scientist into adopting dubious practices in order to publish. Indeed, particularly in recent years, there have been spectacular cases of scientific fraud (see Stroebe et al., 2012 or Fanelli & Tregenza, 2009 for recent discussions). These cases are surely one reason for the current debate surrounding scientific quality management in general and replication in particular. If a scientist reports on an experiment with exciting new results, one way to check if the experiment has actually been carried out and produced the reported results is replication.

There are other indirect ways to check on the authenticity of the experiment, namely, by doing consistency checks or identifying suspicious numeric results (Simonsohn, in press), but these are not always easy to apply. But besides these spectacular, however rare, cases of genuine fraud, even the most honourable and meticulous scientists are in danger of committing methodological errors (see Herndon et al., 2013). Some of these are harmless, while some might utterly distort the results. Particularly, the application of statistical methods provides a plethora of possible pitfalls, and it has often been demonstrated that even well-established scientists may not always fully understand the methods they apply (Kahneman, 2011; Gigerenzer et al., 2004). Again, replication is the most direct way of double-checking experimental results along with the careful examination of the reported experimental procedures and the re-analysis of the original data, if available. Furthermore, even experiments which are conducted without any methodological flaws might still be influenced and distorted by (1) sheer randomness, (2) uncontrollable or hidden confounding factors (Meehl, 1978) and (3) (consciously or unconsciously) applied researcher degrees of freedom (Simmons et al., 2011). The first factor is inevitably present, and statistical methods have been developed to deal with it in a principled way. Hidden factors are quite hard to examine, by the very definition, and researchers tend to assume that they are absent or have no influence on the problem at hand. As long as no contradictions appear or new, unexplainable observations are made, it is in fact reasonable (in the very sense of the word) to assume so. But scientists should be aware of the set of ancillary assumptions and the influence of the specific conditions involved in every experiment (Meehl, 1978). A typical example of an ancillary assumption is the validity and reliability of a standardised test battery used to measure specific traits within a participant sample. The specific conditions comprise all details of the experimental setup: setting, equipment, personality/behaviour of the research assistant, cultural and socio-economical background of the subjects, etc. Hence, even if the

replication of an experiment using an identical test battery fails, this might not fundamentally refute the original hypotheses but just the ancillary assumptions or the specific conditions, which cannot be determined from a single experiment alone. On the other hand, if an experiment provides positive evidence for a theory or hypothesis, this also does not allow a full generalisation to arbitrary populations and experimental settings due to the ancillary assumptions and specific conditions involved (e.g., the WEIRD discussion, Henrich et al., 2010).

Another critical issue concerns researcher degrees of freedom, which are associated with more or less sound and unsound practices (Simmons et al., 2011). To some extent, these are an inevitable and necessary part of everyday scientific praxis. However, Simmons et al. (2011) have suggested guidelines for authors, editors, and reviewers of empirical psychological papers to limit the researcher degrees of freedom and increase transparency in reporting analytical procedures. On the whole, replicating empirical studies is a relatively easy-to-apply yet effective counter-measure to the confounding and distorting factors inherent in scientific research.

Exact and conceptual replication

The question to what extent so-called conceptual replications can fulfil the same functions as exact replications is currently still debated (e.g., Pashler & Harris, 2012 and references therein). First, it must be noted that in psychology, exact replications in the very strict sense cannot exist, since at least location and time of an independently replicated experiment cannot be the same as in the original one. Of course, time and location are not generally the only factors that differ, but often also an unknown number of nearly uncontrollable factors, such as the researchers involved in running the studies,

unrepresentative samples, quirky or badly maintained measurement instruments, rounding errors, non-linear and chaotic effects, subject biases, and so forth (cf. Meehl, 1978 for a discussion). However, these uncontrollable factors are generally assumed to have only small effects and are in practice regarded as experimental noise – whether this is justified in all cases or not is a different matter. Moreover, a replication sample of human participants will come from a, however slightly, different background. There will be differences in details of the research methodology, in the given instructions, in stimuli, in the equipment used, etc. (the specific conditions mentioned above). But, the reproduction of the ancillary assumptions and specific conditions of an experiment as closely as possible is normally considered to be an exact replication. A conceptual replication, on the other hand, might differ in the specific hypotheses deduced from a common theory and possibly in the employed ancillary assumptions. To establish universality for a scientific theory, conceptual replications must be carried out. This follows from the principle of empirical induction and is also necessary in the Popperian concept of falsification (Popper, 1934). For instance, if Newton's law of gravitation had only been formulated for planets—in an attempt to explain Kepler's laws—it would certainly be of much less use than the general version pertaining to every object with mass. Replicating Newton's original experiments by using only planetary motion cannot be used to establish a law that holds for apples as well. At the same time, any theory induced from a particular experiment has to be tested under different conditions. Of course, exact replications of experiments are more likely to detect fraud or methodological errors, but a sufficiently large amount of failed conceptual replications would serve this function as well.

Nevertheless, it seems reasonable to start from an exact replication as a more reliable and effective method. After all, if a long row of conceptual replications ultimately fails to establish the original theory, a large pool of resources may likely have been wasted along the way, which could have been saved if the original experiment had been replicated in the first

place. The same argument implies that replication efforts should be carried out in a hierarchical manner: Seminal and ground-breaking experiments, which might have the power to change a whole set of theories or models, should be replicated first in as similar conditions as possible. Afterwards, a series of conceptual replications and variations should be conducted to establish the extent to which the initial experiment can be generalised and to refine and possibly falsify the induced theory. Finally, subsequent meta-analyses are helpful and inevitable in summarising and interpreting findings from many replications and in establishing a believable assessment of a research topic.

Why is there a lack of replication in music psychology?

Lack of replication and meta-analyses in music psychology

There is a considerable lack of exact replication experiments and meta-analyses in music psychology; however, this is not easy to quantify in precise numbers. As a result of a literature search in early 2013, we were able to identify fewer than 10 meta-analyses. Only the "Mozart effect" (Rauscher et al., 1993) received considerable attention with two meta-analyses. A search for the keyword "replication" and "meta-analysis" in the content of four leading music psychology journals (Musicae Scientiae, Music Perception, Journal of New Music Research, Psychology of Music) followed by a subsequent cross-check of the retrieved abstracts provided an estimate of the proportion of published replication studies and meta-analyses in the field. The results are given in Table 1 and Table 2. The percentage of replication papers in three of the four major music psychological publications is less than 1%, and the mean percentage is 0.4%. Of these 12 replication papers (out of approximately 3500 published papers), 5 deal with the Mozart effect. The percentage of meta-analytical papers in three of the four major music psychological publications is less than 1%. However, because

replication studies and meta-analyses do not always carry indicative terms in the title or keywords, and might be published outside the four core journals screened here, these numbers might underestimate the true proportion of replication studies in music psychology. A similar and more thorough examination in the field of general psychology by Makel et al. (2012) estimated the proportion of genuine replication studies to be 1.1%, which is of the same order of magnitude as our findings in the music psychology domain. Altogether, this readily indicates that genuine replication studies and meta-analyses are very rare in music psychology. It would be reasonable to demand at least one independent replication for each original experiment, which would imply that published replication studies should comprise at least 50% of the music psychology literature—about one order of magnitude higher than the current percentage. This indicates that most research in music psychology is not supported or cross-checked by independent evidence. Hence, the question arises what the reasons for this might be.

==========

insert Table 1 about here

==========

==========

insert Table 2 about here

==========

Costs, benefits and replication

Beliefs serve as the *ultima ratio* in human action planning. But not all actions are of the same importance to an individual or a community. For every non-erratic action, potential

costs and benefits are explicitly or implicitly assigned. The higher the benefits and costs of an action, the more important it becomes that the beliefs preceding this action are reliable and valid, since the risks of making false decisions based on insufficient knowledge should be minimised. Transferring this general idea to the scientific domain, it follows that the more costly the implementation and the greater the possible benefits to be gained from the application of the scientific knowledge, the more reasonable it is to apply greater efforts towards achieving reliable and valid theories. Accepting this line of reasoning, we can distinguish between high/low-gain and high/low-cost sciences. The gains and costs of a science are not fixed entities, but depend on the specific needs and values of the society that provides funding (external gains and cost) and of the research community (internal gains and costs). Together they form a complex interrelated system of benefits and costs, but for the sake of argument we will consider gains here as the external, visible benefits and costs as primarily financial costs. Examples of visible gains are technological outcomes that are widely disseminated and used within society. To become visible in this sense, technologies have to be highly reliable and valid, which can only be achieved on the basis of highly reliable and valid models and theories. In fact, widespread, practical use of a technology also provides constant implicit evidence for the reliability and validity of the underlying scientific theories. Typically, high-gains and high-costs are expected to be correlated since low-gain/high-costs sciences are strongly disadvantaged for obvious reasons, whereas high-gain/low-cost sciences are rare because this combination stimulates very rapid progress which quickly develops into a high-gain/high-cost state. High-gain/high-cost sciences maintain higher standards of reliability and validity by definition, whereas low-gain/low-cost sciences can be expected to employ less strict standards. The reason for this is that the consequences of developing an inadequate or useless theory in low-gain sciences are more tolerable; it is reasonable to minimise costs at the loss of quality. Because science funding bodies do not

expect high-gains from low-gain sciences, research in these disciplines is much less subsided and financially supported. In a constant struggle for resources, this can lead to several disadvantageous consequences. (1) To minimise costs, quality management, particularly replication studies, tend to be neglected. (2) Research is devised for the main purpose of "telling interesting stories," which helps in popularizing and advertising the field and its researchers. The benefits of just having "some ad-hoc theory" that tells a good story are high for researchers in a low-gain science, because it at least satisfies curiosity—a psychological benefit that is a primary driver for conducting and consuming research, regardless of whether the theory is 'true' or not. (3) Research is specifically conducted to demonstrate that the gains from this discipline are actually higher than generally assumed. This is of course a legitimate endeavour, but it risks being biased towards certain desired outcomes. Unfortunately, all of these practices are counter-productive to achieving reliable and valid theories, which in turn hinder the development of real successful technologies and therefore truly visible gains. This puts a scientific discipline in danger of becoming a self-satisfying system and of losing connection to its original field of application and its mission to discover the 'true' principles that underlie the phenomena being studied (thus becoming a game that is played for its own sake, a "Glasperlenspiel", Hesse, 1943/1992).

It might be reasoned that the lack of rigour in low-gain sciences is primarily explained by the sparse availability of financial resources, but we argue that it is due to a combination of low-gain and low-cost factors as well as a lack of inherent risk. Low-gain sciences tend to favour a storytelling mode of research over consolidating, replicative, and, hence at times, dull research, in order to advertise its gains to the external world, while offering at the same time benefits by increasing the entertainment factor. Low-costs, on the other hand, lead to scarce resources and cost minimisation mostly on part of quality management, which in turn

is only possible because the risk of costly consequences from inadequate theories is relatively low.

One might consider the possibility that music psychology is a low-gain/low-cost science in the above defined sense. This possibility is entertained by the thought that, to a large extent, it might not matter to the outside world of musicians, music teachers, and ordinary music listeners whether music psychological theories are true, adequate and useful. This thought is not new to the music psychology community. John Sloboda, for example, expressed a similar idea: "Suppose all the music psychology in the world had never been written and was expunged from the collective memory of the world, as if it had never existed, how would music and musicians be disadvantaged? Would composers compose less good music, would performers cease to perform so well, would those who enjoy listening to it enjoy it any less richly?" (Sloboda, 2005, p. 395-396). Of course, the criticism inherent in this question applies not only to music psychology but to a number of other disciplines as well, including social psychology and areas within cognitive psychology, including cognitive neuroscience. However, there is a divergence here from other research areas, such as material sciences, medical research, pharmacology, genetics, and engineering, including audio engineering and the development of audio codecs. In these areas, inadequate theories and wrong or imprecise model predictions often have considerable financial cost implications or serve to determine the competitiveness of commercial products. In these high-gain sciences, the focus on core and benchmarking experiments and replication of experimental findings is very much built into research infrastructures.

Sloboda's question and the associated ideas represent an extreme hypothetical thought-experiment with the primary aim of provoking a careful scrutiny of motivations and research goals in music psychology. An in-depth discussion of this thought-experiment is beyond the scope of this paper but regardless of what one's answer to Sloboda's thought-provoking

question might be, it seems to be much less controversial to recognise that currently the outside world does not impose strong pressures or very high demands on music psychology to deliver rock-solid theories. This is probably connected to the fact that music itself is not generally regarded as an absolute necessity in life (compared to health, education, poverty, food, housing or technology), though music is a steady companion in most people's lives, a widely traded intangible good, and an important part of one's cultural identity.

The comparatively large number of replication studies focussing on the Mozart effect provides additional indirect evidence along the same lines. A valid and reliable effect of this kind would have moved music psychology into the realm of a high-gain science and consequently spawned high efforts into investigating the truthfulness of the effect. Unfortunately, this did not kick-start a general culture of replication in music psychology.

Taken together, the assumption that music psychology is a low-gain/low-cost science might explain why the urge to maintain high levels of quality is rather weak. It is a relatively small field, where resources in general and for replication in particular, are scarce. Since high quality is not of utmost importance, other values such as novelty, creativity, and originality for the sake of storytelling can outweigh values such as validity and reliability. Researchers cannot gain much prestige and funding from replication experiments, and instead will strive for novel and surprising results. In sum, we argue that the implicit perception of music psychology as a low-cost/low-gain science together with a missing structure for incentivising empirical replications are the main causes for a lack of replication experiments in music psychology.

Discussion and outlook

There is a growing concern about the reliability and validity of empirical knowledge in the social sciences, particularly in psychology (Nosek et al., 2012: Asendorpf et al., 2013). One of the main reasons for this concern is the lack of replication and reproducibility. Other reasons pertain to questionable research practices, researcher degrees of freedom, methodological fallacies, and cognitive and ideological biases. Altogether, this leads to the uneasy question, What do we really know in music psychology? With respect to this scepticism, the question arises, What are the practical consequences for the field?

We see two basic operating modes for a low-gain/low-cost science:

Mode 1: Gathering a larger amount of more or less established facts, theories, and models ("Storytelling-Mode").

Mode 2: Bundling efforts to build a smaller, but firmly believable set of effects, theories, and models ("Conservative-Mode").

However, deciding whether Mode 1 or 2 should take priority might not be an easy decision, and every researcher has to take a very personal stand here. We have argued in this paper that currently music psychology is operating predominantly in Mode 1. Undoubtedly, music psychology as a discipline has matured a great degree over the last 30 years. This maturation process has been supported by the existence of specialised journals with rigorous peer-review protocols, through the publication of several excellent handbooks (e.g. Deutsch, 2013; Hallam et al., 2008) and textbooks (e.g. Thompson, 2008; North & Hargreaves, 2008; Tan et al., 2010; Hodges & Sebald, 2011; Koelsch, 2012), and by the establishment of dedicated music psychology training programmes around the globe. However, we argue that despite the fact that research in music psychology is thriving and the field is constantly expanding, it is still behind in terms of research quality management and the reliability of its theories and models. The lack of established routines for independent replications of widely

cited experiments and important research findings is a major reason why we believe that much of the field of music psychology is currently operating in Mode 1.

The quality of knowledge should not be measured primarily by the immediate and visible gains, but first of all by its applied and theoretical value for the field itself. Well-founded beliefs are an end in themselves, because no one can foresee when and if certain pieces of knowledge might become important for application. This is the notion of fundamental research. Only with a rock-solid base of knowledge can truly successful and visible applications emerge. Due to the fast-growing field of Music Information Retrieval (cf. Downie, 2003 for an overview) and other applied and technological sister disciplines, valid and reliable knowledge from music psychology is currently becoming more and more important, for instance, by enhancing the performance of automatic transcription systems, music search engines and music recommendation systems. This opens up whole new fields of application, which might become even economically interesting.

What are the possible measures that could be taken in order to shift music psychology from Mode 1 to Mode 2? Clearly, conducting large-scale replications of key experiments is one important step toward achieving this. There is a range of options to facilitate replication studies (see Nosek et al., 2012 as well as Asendorpf et al., 2013 for more detailed proposals) that fall in two broad categories:

(1) *Elucidation*. Reminding researchers and teaching students about these issues and starting a broader discussion that might ultimately lead to a greater appreciation of replication efforts. This paper and the current special issue on replication are a first step along this path. However, one should not be overly optimistic that elucidation alone will have ground-breaking and sustainable effects, because the incentives to stay in Mode 1 are not likely to change quickly.

(2) *Starting initiatives*. Special replication initiatives (such as this special issue) that are promoted and supported by institutions might be of even greater impact, and also work toward elucidation. One could think of music psychological societies building special task forces, of music psychological journals and conferences promoting and enforcing replication studies, of funding agencies demanding replication as part of larger projects or setting up special programs, and of private research groups and individuals that take personal interest in replication matters. The field of general psychology has already begun to implement some of these ideas, e.g., the Reproducibility Project (Open Science Collaboration, 2012), the Reproducibility Initiative[2], PsychFileDrawer[3] (an online archive for psychological replication experiments), as well as special issues on replication (Pashler & Wagenmakers, 2012; The Psychologist[4]). One important task, particularly with respect to the scarce resources, would be to identify and agree on core experiments that should be replicated, due to their fundamental importance for the field.

On the positive side, there seems to be a current general consensus that replication is a neglected topic in music psychology, regardless of whether one agrees with all points raised in this paper. Replication seems an interesting, fruitful and much needed approach in music psychology at the moment, as evidenced by the great interest in special symposia at recent conferences. We hope that this will result in making replication not only a "hot topic" and short-lived fad, but that the current interest will lead to an open and honest discussion as to how quality management in music psychology can and should be handled, ultimately leading to a culture of replication within our discipline.

References

Asendorpf, J. B., Conner, M., de Fruyt, F., de Houwer, J., Denissen, J. J. A., Fiedler, K., ... & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108–119.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.

Behne, K.-E., & Wöllner, C. (2011). Seeing or hearing the pianists? A synopsis of an early audiovisual perception experiment and a replication. *Musicae Scientiae, 15*(3), 324–342.

Deutsch, D. (2013). *The psychology of music*, (3rd ed.). San Diego: Elsevier.

Downie, J. S. (2003). Music information retrieval. In B. Cronin (Ed.), *Annual review of information science and technology* (pp. 295–340). Medford, NJ: Information Today.

Eerola, T., Louhivuori, J., & Lebaka, E. (2009). Expectancy in Sami Yoiks revisited: The role of data-driven and schema-driven knowledge in the formation of melodic expectations. *Musicae Scientiae, 13*(2), 231–272.

Eerola, T., & Vuoskoski, J. K. (2012). A review of music and emotion studies: Approaches, emotion models and stimuli. *Music Perception, 30*(3), 307-340.

Fanelli, D., & Tregenza, T. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS ONE*, *4*(5), e5738.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561.

Fredrickson, W. E. (1995). A comparison of perceived musical tension and aesthetic response. *Psychology of Music and Music Education, 23*(1), 81–87.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, *15*, 246–263.

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The SAGE handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks, CA: Sage.

Giordano, B. L., & McAdams, S. (2010). Sound source mechanics and musical timbre perception: Evidence from previous studies. *Music Perception, 28*(2), 155–168.

Hallam, S., Cross, I., & Thaut, M. (Eds.). (2008). *Oxford handbook of music psychology*. Oxford, England: Oxford University Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61–135.

Herndon, T., Ash, M., & Pollin, R. (2013). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Working Paper Series, No. 322*, Political Economy Research Institute, University of Massachusetts Amherst.

Hesse, H. (1943). *Das Glasperlenspiel. Versuch einer Lebensbeschreibung des Magister Ludi Josef Knecht samt Knechts hinterlassenen Schriften.* [The glass bead game.] Zürich: Fretz & Wasmuth.

Hodges, D., & Sebald, D. (2011). *Music in the human experience: An introduction to music psychology*. New York, NY: Routledge.

Howson, C., & Urbach, P. (1989). *Scientific reasoning: The Bayesian approach*. La Salle, IL: Open Court.

Huron, D., & Ollen, J. (2003). Agogic contrast in French and English themes: Further support for Patel and Daniele (2003). *Music Perception, 21*(2), 267–271.

Huron, D., & Sellmer, P. (1992). Critical bands and the spelling of vertical sonorities. *Music Perception, 10*(2), 129–150.

Ivanov, V. K., & Geake, J. G. (2003). The Mozart effect and primary school children. *Psychology of Music, 31*(4), 405–413.

Jackson, C. S., & Tlauka, M. (2004). Route-learning and the Mozart effect. *Psychology of Music, 32*(2), 213–220.

Jaynes, E. T. (2003). *Probability theory: The logic of science.* Cambridge, England: Cambridge University Press.

Juslin, P. N., & Laukka P. (2004). Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research, 33*(3), 217–238.

Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.

Kämpfe, J., Sedlmeier, P., & Renkewitz, F. (2011). The impact of background music on adult listeners: A meta-analysis. *Psychology of Music, 39*(4), 424–448.

Koelsch, S. (2012) *Brain and music.* Hoboken, NJ: Wiley-Blackwell.

Kopiez, R., & Platz, F. (2009). The role of listening expertise, attention, and musical style in the perception of clash of keys. *Music Perception, 26*(4), 321–334.

Livingstone, S. R., & Thompson, W. F. (2006). Multimodal affective interaction: A comment on musical origins. *Music Perception, 24*(1), 89–94.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161–175.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834.

Miksza, P. (2011). Relationships among achievement goal motivation, impulsivity, and the music practice of collegiate brass and woodwind players. *Psychology of Music, 39*(1), 50–67.

Mishra, J. (2013). Improving sightreading accuracy: A meta-analysis. *Psychology of Music, 0*(0), 1–26.

Nesselroade, J. R., Stigler, S. M., & Baltes, P. B. (1980). Regression toward the mean and the study of change. *Psychological Bulletin, 88*(3), 622–637.

North, A., & Hargreaves, D. (Eds.). (2008). *The social and applied psychology of music.* Oxford, England: Oxford University Press.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531–536.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530.

Platz, F., & Kopiez, R. (2012). When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music. *Music Perception, 30*(1), 71–83.

Popper, K. R. (1934). *Logik der Forschung* [The logic of scientific discovery]. Wien, Austria: Julius Springer.

Rauscher, F. H., Shaw, G. L., & Ky, C. N. (1993). Music and spatial task performance. *Nature*, *365*(6447), 611–611.

Rauscher, F. H. (2006). The Mozart effect in rats: Response to Steele. *Music Perception, 23*(5), 447–453.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437–437.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Simonsohn, U. (in press). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science*.

Sloboda, J. A. (2005). *Exploring the musical mind: cognition, emotion, ability, function*. Oxford, England: Oxford University Press.

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*(6), 670–688.

Swanwick, K. (1991). Further research on the musical development sequence. *Psychology of Music and Music Education, 19*(1), 22–32.

Tan, S. L., Pfordresher, P. Q., & Harré, R. (2010). *Psychology of music: From sound to significance.* London, England: Routledge and Psychology Press.

Thompson, W. F. (2008). *Music, thought, and feeling: Understanding the psychology of music.* New York, NY: Oxford University Press.

Footnotes

[1]http://cds.cern.ch/record/1165534/files/CERN-Brochure-2009-003-Eng.pdf [21.4.2013]

[2] https://www.scienceexchange.com/reproducibility[3.3.2013]

[3] http://www.psychfiledrawer.org/ [3.3.2013]

[4] http://tinyurl.com/cogdy5x[3.3.2013]

Table captions

Table 1. Numbers and proportions of replication papers (search string “replication”) in four major music psychology journals.

Table 2. Numbers and proportions of meta-analyses papers (search string “meta-analysis”) in four major music psychology journals.

Table 1

Journal	Replication Papers	Mozart Effect papers	Total No. of Papers	Proportion of Replication Papers	References
Musicae Scientiae	2	0	~190	~1%	Eerola et al. (2009), Behne & Wöllner (2011)
Music Perception	5	2	~2000	~0.25%	Huron & Ollen (2003), Huron & Sellmer (1992), Kopiez & Platz (2009), Rauscher (2003), Rentfrow et al. (2012)
Journal of New Music Research	0	0	~540	~0%	
Psychology of Music	5	2	~800	~0.5%	Fredrickson (1995), Ivanov & Geake (2003), Miksza (2011), Jackson & Tlauka (2004), Swanwick (1991)
Total	12	5	~3530	~0.4%	

Table 2

Journal	Meta-analysis Papers	Mozart Effect papers	Total No. of Papers	Proportion of Replication Papers	References
Musicae Scientiae	0	0	~190	~0%	
Music Perception	4	0	~2000	~0.2%	Eerola & Vuoskoski (2012), Giordano & McAdams (2010), Livingstone & Thompson (2006), Platz & Kopiez (2012),
Journal of New Music Research	1	0	~540	~0.2%	Juslin & Laukka (2004)
Psychology of Music	2	0	~800	~0.3%	Kämpfe et al. (2011), Mishra (2013).
Total	7	0	~3530	~0.2%	