A REVIEW OF VALIDITY AND ITS RELATIONSHIP TO MUSIC INFORMATION RESEARCH

Bob L. T. Sturm
Division of Speech, Music and Hearing
KTH Stockholm, Sweden
bobs@kth.se

Arthur Flexer
Institute of Computational Perception
Johannes Kepler University Linz, Austria
arthur.flexer@jku.at

ABSTRACT

Validity is the truth of an inference made from evidence and is a central concern in scientific work. Given the maturity of the domain of music information research (MIR), validity in our opinion should be discussed and considered much more than it has been so far. Puzzling MIR phenomena like adversarial attacks, horses, and performance glass ceilings become less mysterious through the lens of validity. In this paper, we review the subject of validity as presented in a key reference of causal inference: Shadish et al., *Experimental and Quasi-experimental Designs for Generalised Causal Inference* [1]. We discuss the four types of validity and threats to each one. We consider them in relationship to MIR experiments grounded with a practical demonstration using a typical MIR experiment.

1. INTRODUCTION

The multi-disciplinary field of Music Information Research (MIR) is focused on making music and information about music accessible to a variety of users. This ranges from systems for search and retrieval, to recommendation, and even to more creative applications like music generation. The effectiveness and reliability of MIR systems are of prime importance to the MIR researcher, not to mention other stakeholders. The researcher thus performs experiments to compare approaches for modeling and retrieving music data. A principal focus is on users, but the cost of performing experiments with users is high, and the replicability of such studies is difficult. This has motivated the *Cranfield Paradigm* [2]: computer-based experiments where "test collections" serve as proxies for human users. While such an approach is inexpensive and replicable, its relevance and reliability for MIR, and information retrieval in general, have been questioned [3, 4].

Under the Cranfield Paradigm, state-of-the-art MIR systems perform exceptionally well in reproducing the ground truth of some datasets, e.g., inferring rhythm, genre or emotion from audio data. This leads to conclusions that the systems are actually learning to perform the task believed necessary to recover the ground truth from audio data. However, slight and irrelevant transformations of the audio, e.g., "adversarial attacks", can suddenly render these systems ineffectual [5–9]. Such attacks can reveal what an MIR system is relying on for its success. In one case, a "genre recognition" system relies on infrasonic signatures that are imperceptible and irrelevant for human listeners [8]. In another, a "rhythm recognition" system is recognising tempo instead of rhythm, a confounding originating from the data collection [6]. Systems relying on such "tricks" have been called "horses" [5]. A related topic in MIR is "glass ceilings" [10, 11], i.e., that an observed barrier to improving system performance to perfect or human level is claimed as coming from psychophysical and cultural factors of music missing from features extracted from audio recordings [12].

In order to better understand the problems described above it is necessary to consider what lies at the heart of any experiment: the relationship between conclusions drawn from its results and their *validity*, or "truth value" [1]. Ideally, an experiment will be carefully designed and implemented to answer a well-defined hypothesis. Its components – units, treatments, design, observations, and settings – should be carefully operationalised (translated from theory into practice) to maximize quality and minimize cost (e.g., money and time). This is the purview of the discipline *Design of Experiments*: how can one get the strongest evidence for the least cost?

Despite a small chorus of calls to improve validity of conclusions in MIR, e.g., [4–6, 13–19], there has yet to be published a systematic and critical engagement of what validity means in the context of MIR, and how to consider it when designing, implementing and analyzing experiments. In this paper, we focus on the four principal types of validity in Shadish et al. [1], an authoritative resource about validity in causal inference and experimental science. Other typologies exist, e.g., [20], but we use that of Shadish et al. [1] because it is an established point of reference, and has already been mentioned in the context of MIR, e.g., in [4]. We review the four types of validity and present actionable questions that can help MIR researchers to scrutinize the conclusions they draw from their experiments. We ground our general discussion of validity in this paper by a practical demonstration, [1] which presents a typical MIR experiment

[1] See supplementary material here: https://github.com/boblsturm/mirvalidititytutorial

Model	Accuracy	Precision	Recall	f1-score
LDA	0.714	0.711	0.711	0.703
QDA	0.719	0.715	0.723	0.717
1NN	0.662	0.644	0.635	0.638
3NN	0.681	0.673	0.651	0.656
5NN	0.719	0.699	0.687	0.689
7NN	0.695	0.669	0.656	0.659
9NN	0.700	0.681	0.664	0.668
unif	0.12 ± 0.02	0.13 ± 0.03	0.12 ± 0.02	0.12 ± 0.02
freq	0.13 ± 0.02	0.13 ± 0.03	0.13 ± 0.02	0.13 ± 0.02
maj	0.16	0.02	0.12	0.03

Table 1. Accuracy, and macro-averaged precision, recall and f1-score observed for several models in a testing partition of BALLROOM [21]. The performance of two models selecting labels randomly (with standard deviation) are shown in the rows labeled: *unif* samples labels uniformly; *freq* samples labels according to training data label frequency. The last row *maj* shows the performance of a model choosing the label most frequent in the training data.

that exemplifies a considerable amount of MIR research: music classification using machine learning (ML) and a benchmark dataset. We use the BALLROOM dataset [21], which has appeared in dozens of studies seeking to build MIR systems sensitive to rhythm [6]. We partition the dataset into training and testing sets, extract features and train ML models, then label test set recordings and count coincident ground truth labels, and finally compute figures of merit for the different ML models. Table 1 presents the results from which we wish to draw valid conclusions.

A less abridged version of this paper [22] integrates the supplementary material in more detail. We hope that these materials will help MIR researchers to design, implement and analyze experiments in MIR and draw valid conclusions, but also convince them that the language of validity is reason. Creative thinking is necessary when examining the truth value of any conclusion drawn from an experiment.

2. COMPONENTS OF EXPERIMENTS

Before discussing the validity of conclusions drawn from an experiment, we must identify its components: units, treatments, design, observations, and settings. *Treatments* are the things applied to units in order to cause an effect (or not in the case of a *control*), *units* are the things that are treated, and *observations* are what is measured on a unit. The *design* specifies which treatment is applied to which unit, and *settings* involve time, place, and condition. To make this more concrete, consider a medical experiment in which the effect of a treatment on blood pressure is being studied. A number of people are sampled from a population, some of whom will receive the treatment while the others receive a placebo (control). The design describes which people get the treatment, and which do not. The observation is the blood pressure of a person after one month. The setting can include particulars of the population (rural or urban), place of treatment (hospital or home), and so on. The experimentalist contrasts blood pressure observations across groups to conclude, e.g., the effect of the treatment (causes a decrease in blood pressure).

Our typical MIR experiment measures the effectiveness of different ML models in predicting the labels of a test recording dataset. There are two ways to see its components. We can see the treatments as the ten models and the units as replicates of the entire testing dataset, or we can see the entire testing dataset as the one treatment and the units as the ten models. Since Table 1 reports figures of merit (observations) of each model on the entire test dataset, the latter interpretation motivates conclusions about the effectiveness of particular models. In this case, the design is simple: each unit (ML model) is given the same treatment (dataset). The setting involves the dataset partitioning, the extracted features, random seeds, software libraries, etc.

3. STATISTICAL CONCLUSION VALIDITY

Statistical conclusion validity is "the validity of inferences about covariation between two variables" [1]. This includes concluding that a covariation exists, and perhaps its strength as well. This is the level at which one is concerned with *statistical significance*, i.e., that an observed covariation between treatment and effect is not likely to arise by chance. As a concrete example, an experiment measuring the effects of two different medicines on lowering blood pressure seeks to determine which of the medicines has the greatest effect, if at all. The statistical conclusion validity of a conclusion resulting from this experiment relies on its power, but can be threatened in other ways. Shadish et al. [1] (p. 45) includes a table of nine different threats to statistical conclusion validity. Four threats relevant to computer-based experiments are: violated assumptions about the statistics underlying the observations (and the use of the wrong statistical test, a *type III error* [23]); a sample size too small to reliably detect covariation (lack of power); the purposeful search for significant results by trying multiple analyses and data selections ("p-hacking" [24]); and increased variance in observations due to the heterogeneity of units.

Are my results statistically significant? Null hypothesis statistical testing (NHST) quantifies whether the observed effects of the treatments on the responses arise by mere chance, as well as the direction of effect and its size. This answers the question: are the results statistically significant? Fundamentals about statistical testing in MIR have already been discussed [25], also for Artificial Intelligence in general [26], and for ML [27]. One must take care in selecting a statistical test to use; each one makes strong assumptions that could be violated. NHST is most straightforwardly applicable to completely randomized experimental designs [28], thereby reducing the possibility of structure in units and treatments interfering with the responses (which results in confounding). Most MIR experiments cannot use complete randomisation because the target population from which samples come is unclear (what is a random sample of "sad" music, with the term "sad" being quite ill-defined?), and so the kinds of conclusions that can be made with NHST in MIR are limited. [2]

[2] Experimental designs that cannot be completely randomised are called *quasi-experimental designs*, another major topic of Shadish et al. [1].

Is the observed statistical significance relevant for a user? In MIR, even if one finds statistical significance, this may not generalise to a perceivable difference for actual users interacting with the "improved" MIR system. As an example from MIR, a crowd-sourced user evaluation [29] demonstrates that there is an upper bound of user satisfaction with music recommendation systems of about 80%, since this was the highest percentage of users agreeing that two systems "are equally good." In addition, for the MIREX task of *Audio Music Similarity and Retrieval* it has been demonstrated [29] that statistically significant differences between algorithms can be so small that they make no practical difference for users.

Let us now consider the typical MIR experiment and reason about what conclusions we can draw from it that have statistical conclusion validity. Table 1 clearly shows that each response of model to the dataset is greater than the random approaches *unif*, *freq* and *maj*. How likely is it that any of the responses of models is due to chance, i.e., that any of the models is actually no better than one of the random approaches? Since we have the empirical distributions for *unif* and *freq*, we can estimate the probability of either of them resulting in, e.g., a macro-average recall at least as large as 0.6: $p < e^{-200}$.[3] Hence, a valid statistical conclusion is that we observe a significant covariation between the use of a machine learning model with these particular features and the responses measured on a specific partition of BALLROOM.

One might consider statistical conclusions relating to the type of ML, i.e., Gaussian modeling (LDA and QDA) vs. nearest neighbour modeling (KNN), or LDA vs. QDA. If we conclude from Table 1 that Gaussian modeling performs better than nearest neighbour modeling with these features on 70/30 partitions of BALLROOM, we would be wrong. This is a "type I error", which is concluding there to be a significant difference when in fact there is none. When we perform this experiment 1000 times with random 70/30 partitions we observe that the difference between the best response of a Gaussian model and the best response of a nearest neighbour model is distributed Gaussian, and that the probability of observing zero difference or less is $p > 0.41$ for any of the figures of merit.

The most general statistical conclusion we can make from Table 1 is that the responses we observe from ML models are highly inconsistent with the responses of choosing randomly. Each ML model knows *something* about BALLROOM linking the features computed from a music recording with its ground truth label. Because we do not know the amount of variation in any response due to partitioning, we cannot make any valid statistical conclusion about which type of ML model is the best for this particular dataset. In order to go further, we must run the experiment multiple times to obtain distributions of the contrasts. Even then, however, we cannot say anything about the *cause* of significant differences yet. This is where the notion of internal validity becomes relevant.

[3] See the supplementary material for an explanation.

4. INTERNAL VALIDITY

Internal validity is "the validity of inferences about whether the observed covariation between two variables is causal" [1]. While statistical conclusion validity is concerned only with the strength of covariation between treatment and responses, internal validity is focused on the *cause* of a particular response to the treatment. Shadish et al. [1] (p. 55) includes a table of nine different threats to causal conclusions. Several of these involve *confounding*, which is the confusion of the treatment with other factors arising from poor operationalisation in an experiment. As a concrete example, consider an experiment measuring the effects of two different medicines on lowering blood pressure, but where one medicine is given to young patients and the other is given to elderly patients. This experimental design confounds the two medicines and patient age, and so the effects caused by the two factors cannot be disambiguated. Any conclusion from this experiment about the effects of the medicines lacks internal validity.

Does my data collection introduce confounds? One's methodology for collecting data might unintentionally introduce structure. For instance, it has been discussed that BALLROOM was assembled by downloading excerpts of music CDs sold at a website selling music for ballroom dance competitions [6]. Ballroom dance competitions are regulated by organisations, e.g., World DanceSport Federation (WDSF),[4] to ensure uniformity of events for competitors around the world. These organisations set strict requirements of tempo of each dance such that high skill is required of the dancers. Hence, the labels of BALLROOM can reflect any of the following: 1) the rhythm of the music; 2) the type of dance performed to the music; 3) the strict tempo requirements of the dance in the context of competition. As a result, good performance in BALLROOM can be due to rhythm detection and/or tempo estimation. Tempo and rhythm are related musical characteristics, but they are not the same thing [30].

Does my data partitioning introduce confounds? Dataset partitioning can also introduce confounds, e.g., "bleeding ground truth." An example is to first segment recordings into short (e.g., 40ms) time frames and then partition these frames into training and testing sets, thus spreading highly correlated features across these sets. In the context of audio-based genre classification, the presence of songs from the same artists or albums in both training and test data has been shown to artificially inflate performance [31, 32]. Audio-based genre classification using very direct representations of spectral content has been shown [33] to degrade more when employing artist/album filters than classification based on more abstract kind of features like rhythmic content (fluctuation patterns). This insight that problems of data partitioning can affect MIR systems in quite different ways and hence change performance rankings has been confirmed in another meta-study [34].

Returning to our typical MIR experiment, of interest is *what* it is in our trained ML models causing their response to be inconsistent with random selection. Knowing how

[4] https://www.worlddancesport.org/

Gaussian models used in LDA and QDA are built – mean and covariance parameters are estimated from training data – an internally valid conclusion is that these models work well in BALLROOM because likelihood distributions estimated from the training data also fit the testing data well. Another internally valid conclusion is that the high performances of these ML models in BALLROOM are caused by the features together with the expressivity of the models capturing information related to the labels in BALLROOM.

With reference to the aims of MIR research, we want to conclude something more specific, e.g., our ML models have learned to recognize the rhythms in BALLROOM. This is certainly one explanation consistent with our observations, but is it the only one? The internal validity of this conclusion relies on a key assumption: inferring the labels of BALLROOM can *only* be the result of learning to discriminate between and identify its rhythms. In other words, we must assume that there is no other way to infer labels in BALLROOM than by perceiving rhythm.

Since we know tempo is highly correlated with rhythm in BALLROOM, we thus perform an experiment to test the sensitivity of our trained ML models to tempo: we alter all test recordings by some amount of pitch-preserving time dilation, and then measure the responses of the models to these new treatments. We see that the responses of all ML models decay to being not significantly different from random selection with dilations in the range of ± 15 %. We see this intervention clearly reveals the extent to which the ML models we test rely on the tempi in the test data.

The experimental design of the typical MIR experiment does not account for the structure present in the dataset; we do not control for other ways of inferring the labels of BALLROOM, which are guaranteed to exist by its very construction. From Table 1 and our experimental design, we thus cannot be any more specific in our causal inference than this: the responses of our ML models are caused by their having learned *something* about BALLROOM. This then calls into question how comparing predictions with ground truth in BALLROOM relates to the ability we might actually want to measure, that is the recognition of rhythm. This is where the notion of construct validity becomes relevant.

5. CONSTRUCT VALIDITY

Construct validity is "the validity of inferences about the higher order constructs that represent sampling particulars" [1]. This involves the relationship between what is meant to be inferred by the experimentalist from an experiment and what is actually measured, i.e., the *operationalisation* of the experimentalist's intention. For instance, directly measuring the blood pressure of a person involves an invasive procedure inserting a measuring device in their veins. Blood pressure can be measured less invasively but indirectly by externally applying known pressure to a vein and listening for when blood flow ceases. Knowledge about the incompressibility of liquids in closed systems makes the measurement of pressure in the balloon a relevant measure of blood pressure. Shadish et al. [1] (p. 73) includes a table of fourteen different threats to construct validity, but several of these are irrelevant to computer-based experiments. The main threat is a questionable relationship between what is being measured and what is intended to be measured. Selecting a measure by convenience but not relevance, sampling from convenient populations, and a lack of definition of what is intended to be measured, are threats to construct validity. Construct validity involves more than just how something is measured; it also involves what is measured and in what settings.

How is classification accuracy, or any figure of merit, in a labeled music dataset related to X*?* Two examples in MIR are the use of "genre" classification accuracy as an indirect measure of music similarity [11], or user satisfaction (see, e.g., [14] for a discussion). The relationship between these is very tenuous, especially so considering that accuracy itself is an unreliable measure of whether or not a system has learned anything relevant to music [5, 15]. A key reference in this respect is that of Pfungst [35] describing a series of experiments in trying to reliably measure the arithmetic acumen of a horse that was only able to tap out answers. Counting the number of correct answers tapped out by the horse, no matter how many questions are asked, is irrelevant without considering how each question is posed (the setting). The key to Pfungst discovering the cause of the horse's apparent arithmetic acumen involved changing the setting: the questions remained the same, and accuracy of correct response was measured, but how the questions were posed was changed in order to control for different factors of the experiment. The same is true for MIR.

What is the "use case" of the system to be tested? To counter threats to construct validity the MIR experimentalist must operationalise as much as possible the use case of the system to be built and tested. One attempt to do so for music description [36] emphasises the need to define success criteria. The experimentalist must determine how their method of measurement relates to the success criteria, e.g., relating accuracy in genre classification to the satisfaction of a specific type of user.

How can we test the construct validity of a conclusion? One possibility is to assess the outcomes of different experiments which are supposed to measure the same higher order constructs. An example in MIR is to study correlations of different genre classifiers when given identical inputs [18]. Low correlations between classifiers point to problems of construct validity. A related topic is that of adversarial examples, which casts doubt on the conclusion that the high accuracy of an MIR system in some dataset reflects its "perception" of the music in the waveform. Adversarial examples have first been described in image analysis [37], where imperceptible perturbations of input data significantly degraded classification accuracy.

Returning to our typical MIR experiment, we are interested in making construct inferences around the latent ability of rhythm recognition we are supposedly measuring in our ML models. For instance, one construct inference is that our features measure relevant aspects of rhythm in recorded music. In some sense, by their definition from basic signal processing components, our features come from

temporal aspects that are certainly relevant to rhythm. Our features are also reliant on acoustic information, and in particular there being high-contrast differences in onsets captured by spectral flux – hence limiting their relationship to rhythms played by particular kinds of instruments with sharp attacks. However, we have seen above that the features are also indicative of tempo, and that tempo is another path an ML model can use to infer the rhythm label. Hence we are left to question the relationship of our features to the concept we are trying to operationalise, i.e., rhythm.

Having a system label any partition of the BALLROOM dataset provides no reliable measure of a system's ability to recognise rhythm without changing the setting to control for other factors. It is not as simple as choosing a different feature, measure, cross-validation method, or using a particular statistical test. One must change the experiment itself such that *rhythm recognition* is what is actually being measured. This means that BALLROOM can still be useful to measuring the rhythm recognition of an ML model. Indeed, in the previous section we used it to disprove the causal claim that the good performance of the ML systems of Table 1 is caused by their ability to recognize rhythm. Might performance in BALLROOM also be an indication of performance in other datasets focused on rhythm? This is where the notion of external validity becomes relevant.

6. EXTERNAL VALIDITY

External validity is "the validity of inferences about the extent to which a causal relationship holds over variations in experimental units, settings, treatment variables and measurement variables" [1]. More generally, external validity is the truth of a generalised causal inference drawn from an experiment. An example is inferring that medicine found to lower blood pressure in patients living in Germany will also lower blood pressure in people living in Mexico – a conclusion that can lack validity due to differences in diet, living and working conditions, and so on. Another example is that increasing the dose of the medicine will cause blood pressure to lower further in the studied population. If a causal inference we draw from an experiment lacks internal validity, then generalising that inference to include variations not tested will not have external validity. Shadish et al. [1] (p. 87) includes a table of five different threats to external validity, which are in addition to the threats to internal validity. The main threat is that variation of the components of the experiment might destroy the causal inference that holds in the experiment. For instance, a medication may work for the type of illness tested, but that type of illness may not be generalisable to other closely related illnesses.

Does my model generalize to out-of-sample data? The standard approach in evaluating MIR classification systems is to use separate train and test sets in cross-validation experiments to obtain seemingly unbiased estimates of performance. However, if such MIR systems are exposed to independent out-of-sample data often severe loss of performance is observed. One example are experiments on genre recognition where accuracy results do not hold when evaluated across different collections that are not part of the training sets [38, 39]. The results do not generalize to supposedly identical genre labels in different collections, which reflects a lack of external validity. Genre labels like 'Rock' will be used differently by different annotators working on these collections – which is also a threat to construct validity. Another example are how different audio encodings affect subsequent computation of descriptors and classification results [40], or how in general differences in software implementations diminish replicability [41].

Do different raters agree on a ground truth? Human perception of music is highly subjective resulting in possible low inter-rater agreement. Therefore only a certain amount of agreement can be expected if several human subjects are asked to rate the same song pairs according to their perceived similarity, depending on a number of subjective factors [14, 42] like personal taste, listening history, familiarity with the music, current mood, etc. Concerning annotation of music, it has been shown [43] that the performance of humans classifying songs into 19 genres ranges from modest 26% to 71%. Audio-based grounding of everyday musical terms shows the same problematic results [44]. It has even been argued [12] that no such thing as an immovable 'ground' exists in the context of music, because music itself is subjective, highly context-dependent and dynamic.

The lack of inter-rater agreement presents a problem of external validity because inferences from the experiment do not generalize from users or annotators in the experiment to the intended target population of arbitrary users/annotators. It is also a problem of reliability, since different groups of users or annotators with their differing subjective opinions will impede repeatability of experimental results. This lack of inter-rater agreement presents an upper bound for MIR approaches, since it is not meaningful to have computational models going beyond the level of human agreement. Such upper bounds have been reported [14,42,45] for the MIREX tasks of 'Audio Music Similarity and Retrieval' (AMS) and 'Music Structural Segmentation' (MSS). For AMS the upper bound has already been reached in 2009, while for MSS the upper bound is within reach for at least some genres of music. Comparable results exist concerning music structure analysis [46] and chord estimation [47, 48].

Do raters agree with themselves at different points in time? Going beyond the question of whether different annotators agree on a ground truth one can also access what the level of agreement within one person is when faced with identical annotation tasks at different points in time. A high intra-rater agreement would help to overcome the problem of upper bounds in MIR systems since it would make personalization of models meaningful, i.e. to have separate models for individual persons. However, at least for the task of general music similarity it has been shown that intra-rater agreement is only slightly higher than inter-rater agreement [19], with the absolute level also depending on music material and mood of raters at test time. An approach to personalize chord labels for individual annotators via deep learning was more successful [49].

Retuning to the typical MIR experiment, we cannot

	Accuracy	Precision	Recall	f1-score
LDA	0.659	0.647	0.643	0.643
QDA	0.682	0.678	0.672	0.673
1NN	0.622	0.616	0.602	0.604
3NN	0.636	0.629	0.610	0.613
5NN	0.644	0.643	0.617	0.619
7NN	0.647	0.646	0.619	0.621
9NN	0.645	0.643	0.615	0.618
unif	0.12 ± 0.01	0.13 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
freq	0.13 ± 0.01	0.13 ± 0.01	0.12 ± 0.01	0.12 ± 0.01
maj	0.13	0.02	0.12	0.03

Table 2. As in Table 1, models trained in BALLROOM and tested in all of X-BALLROOM [50].

validly conclude that any of our models is recognizing rhythm in general because we do not know if they are recognizing rhythm in BALLROOM. Our dilation intervention experiment in Sec. 4 reveals that all of the models lose their supposed ability to recognize rhythm in BALLROOM, so there is no reason to infer they will recognize rhythm elsewhere. One causal conclusion we might make is that our models perform well in BALLROOM because they have learned something about BALLROOM – a curated set of recordings downloaded from a specific website in 2004. Might they have learned something about other recordings from that same website, but collected many years later?

The extended BALLROOM dataset (X-BALLROOM) [50] consists of 3,484 audio recordings in the same eight dance styles or music rhythms as BALLROOM, but downloaded from the same website over a decade later. This gives us a chance to test our conclusion. The figures of merit measured from our models trained in BALLROOM but applied to all of X-BALLROOM are shown in Table 2. We still see significant covariation between response and the use of ML with our features. By and large, whatever concepts our ML models have learned about BALLROOM carry over to X-BALLROOM – but we still do not know whether or not those concepts have to do with rhythm.

7. CONCLUSION

This paper provides a review of the notion of validity based on the typology given in Shadish et al. [1]. It brings together the few sources in MIR that mention validity, and several sources that do not but are related. This paper does not aim to prescribe how to design and perform experiments such that valid conclusions can be drawn from them. Instead, it aims to bring within the realm of MIR what validity means, why it is important, and how it can be threatened. One thing to reiterate is that one does not talk about the "validity of an experiment". An experiment does not possess "truth value". Validity is a property of a conclusion made given evidence collected from an experiment. The components of an experiment – units, treatments, design, observations, and setting – have major consequences for the validity of conclusions drawn from it, whether it is statistical conclusion validity, internal validity, construct validity, or external validity.

In MIR the predominant experimental methodology is the Cranfield Paradigm: train a model on a partition of a dataset and count the number of correct answers on another partition. This kind of experiment is inexpensive, and provides numbers that can be compared in ways that convince peer reviewers that progress has been accomplished [51]. Despite various appeals [14, 52] and beseechings [4, 5, 15, 16, 19, 29, 42], such an experimental approach is still standard in the field and its serious flaws are ignored. Any conclusion from this experiment that is more general than "the system has learned something about the dataset" lacks internal, construct and external validity. This does not mean that all such inferences are false, just that they cannot follow from the experiment as designed and implemented. Reproducing the ground truth of a dataset represents a beginning and must be followed by a search for the causes of the observed behavior. One must resist the urge to conclude that an MIR system must be doing whatever is hoped for.

Shadish et al. [1] provides an established starting point for MIR, but there exist other types of validity. For instance, Lund [20] revises the typology of [1] to address ambiguities between causes and treatments, to better define aspects of settings, and to establish a hierarchical ordering of five types of validity: statistical conclusion, causal, construct, generalization and theoretical. An important distinction in this typology is its emphasis on a major aim of basic research: to contribute theory. Other kinds of validity include ecological, convergent, and criterion [13], but these still deal with the kind of conclusion one is drawing from evidence collected in some way.

As a final note, a frustration when encountering Shadish et al. [1] as an engineer is that of its 623 pages there are only five pages with at least one equation on them. Instead, Shadish et al. [1] describe experiments and how each type of validity manifests in the conclusions drawn, with specific threats to the reasoning of those conclusions. Experiments, not to mention experimentalists, are such complex assemblages that expressing them in formal ways that appear to permit the computation of numbers that relate to each type of validity would probably have very limited applicability, and then only be understood by a limited audience. The language of validity is *reason*, and we hope this article will inspire MIR researchers to think creatively about the phenomena they observe to discover their causes.

8. ACKNOWLEDGMENTS

We thank J. Urbano and H. Maruri-Aguilar for helpful discussions during the drafting of previous versions of this paper, as well as the constructive criticisms of reviewers from the Transactions of the Society for Music Information Retrieval. The contribution of Sturm is supported by a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 864189 MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism). The contribution of Flexer is supported by funding from the Austrian Science Fund (FWF, project numbers P 31988 and P 36653). For the purpose of open access, the authors have applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission.

9. REFERENCES

[1] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2002.

[2] C. W. Cleverdon, "The significance of the Cranfield tests on index languages," in *Proc. Int. ACM SIGIR Conf. Research and Development in Info. Retrieval*, 1991, pp. 3–12.

[3] E. M. Voorhees, "The philosophy of information retrieval evaluation," in *Proc. Cross-Language Evaluation Forum*, 2001.

[4] J. Urbano, M. Schedl, and X. Serra, "Evaluation in music information retrieval," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 345–369, 2013.

[5] B. L. Sturm, "A simple method to determine if a music information retrieval system is a 'horse'," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1636–1644, 2014.

[6] ——, "The "horse" inside: seeking causes behind the behaviors of music content analysis systems," *Computers in Entertainment (CIE)*, vol. 14, no. 2, pp. 1–32, 2017.

[7] C. Kereliuk, B. L. Sturm, and J. Larsen, "Deep learning and music adversaries," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2059–2071, 2015.

[8] F. Rodríguez-Algarra, B. L. Sturm, and H. Maruri-Aguilar, "Analysing scattering-based music content analysis systems: Where's the music?" in *Proc. Int. Symp. Music Information Retrieval*, 2016, pp. 344–350.

[9] K. Prinz, A. Flexer, and G. Widmer, "On end-to-end white-box adversarial attacks in music information retrieval," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, pp. 93–104, 2021.

[10] J.-J. Aucouturier and F. Pachet, "Improving timbre similarity: How high is the sky?" *J. Neg. Results Speech Audio Sci.*, vol. 1, no. 1, pp. 1–13, 2004.

[11] T. Pohle, E. Pampalk, and G. Widmer, "Evaluation of frequently used audio features for classification of music into perceptual categories," in *Proc. Int. Workshop Content-Based Multimedia Indexing*, 2008.

[12] G. A. Wiggins, "Semantic gap?? schemantic schmap!! methodological considerations in the scientific study of music," in *Proc. Int. Symp. Multimedia*. IEEE, 2009, pp. 477–482.

[13] J. Urbano, "Information retrieval meta-evaluation: Challenges and opportunities in the music domain," in *Proc. Int. Symp. Music Information Retrieval*, 2011, pp. 609–614.

[14] M. Schedl, A. Flexer, and J. Urbano, "The neglected user in music information retrieval research," *Journal of Intelligent Information Systems*, vol. 41, no. 3, pp. 523–539, 2013.

[15] B. L. Sturm, "Classification accuracy is not enough: On the evaluation of music genre recognition systems," *J. Intell. Info. Systems*, vol. 41, no. 3, pp. 371–406, 2013.

[16] ——, "Revisiting priorities: Improving MIR evaluation practices," in *Proc. ISMIR*, 2016.

[17] J. Urbano and A. Flexer, "Statistical analysis of results in music information retrieval: why and how (abstract)," in *Proc. Int. Symp. Music Information Retrieval*, 2018, pp. xli–xlii.

[18] C. C. Liem and C. Mostert, "Can't trust the feeling? how open data reveals unexpected behavior of high-level music descriptors," in *Proc. Int. Symp. Music Information Retrieval*, 2020, pp. 240–247.

[19] A. Flexer, T. Lallai, and K. Rašl, "On evaluation of inter- and intra-rater agreement in music recommendation," *Transactions of the International Society for Music Information Retrieval*, vol. 4, no. 1, pp. 182–194, 2021.

[20] T. Lund, "A revision of the campbellian validity system," *Scandinavian J. Educational Research*, vol. 65, no. 3, pp. 523–535, 2021.

[21] S. Dixon, F. Gouyon, and G. Widmer, "Towards characterisation of music via rhythmic patterns," in *Proc. Int. Symp. Music Information Retrieval*, 2004, pp. 509–517.

[22] B. L. T. Sturm and A. Flexer, "Validity in music information research experiments," *arxiv*, vol. arXiv:2301.01578, 2023.

[23] A. W. Kimball, "Errors of the third kind in statistical consulting," *J. American Statistical Assoc.*, vol. 52, no. 278, pp. 133–142, June 1957.

[24] M. L. Head, L. Holman, R. Lanfear, A. T. Kahn, and M. D. Jennions, "The extent and consequences of p-hacking in science," *PLOS Biology*, vol. 13, no. 3, pp. 1–15, 2015.

[25] A. Flexer, "Statistical evaluation of music information retrieval experiments," *Journal of New Music Research*, vol. 35, no. 2, pp. 113–120, 2006.

[26] P. R. Cohen, *Empirical methods for artificial intelligence*. MIT press Cambridge, MA, 1995, vol. 139.

[27] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. New York, NY, USA: Cambridge University Press, 2011.

[28] R. A. Bailey, *Design of comparative experiments*. Cambridge University Press, 2008.

[29] J. Urbano, J. S. Downie, B. Mcfee, and M. Schedl, "How significant is statistically significant? the case of audio music similarity and retrieval." in *Proc. Int. Symp. Music Information Retrieval*, 2012, pp. 181–186.

[30] W. A. Sethares, *Rhythm and Transforms*. Springer, 2007.

[31] E. Pampalk, A. Flexer, G. Widmer *et al.*, "Improvements of audio-based music similarity and genre classificaton." in *Proc. Int. Symp. Music Information Retrieval*, 2005, pp. 634–637.

[32] A. Flexer and D. Schnitzer, "Effects of album and artist filters in audio similarity computed for very large music databases," *Computer Music Journal*, vol. 34, no. 3, pp. 20–28, 2010.

[33] A. Flexer, "A closer look on artist filters for musical genre classification," in *Proc. Int. Symp. Music Information Retrieval*, 2007, pp. 341–344.

[34] B. L. Sturm, "The state of the art ten years after a state of the art: Future research in music information retrieval," *Journal of New Music Research*, vol. 43, no. 2, pp. 147–172, 2014.

[35] O. Pfungst, *Clever Hans (The horse of Mr. Von Osten): A contribution to experimental animal and human psychology*. New York: Henry Holt, 1911.

[36] B. L. Sturm, R. Bardeli, T. Langlois, and V. Emiya, "Formalizing the problem of music description," in *Proc. Int. Symp. Music Information Retrieval*, 2014, pp. 89–94.

[37] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *Proc. Int. Conf. Learning Representations*, 2014.

[38] D. Bogdanov, A. Porter, H. Boyer, X. Serra *et al.*, "Cross-collection evaluation for music classification tasks," in *Proc. Int. Symp. Music Information Retrieval*, 2016, pp. 379 – 385.

[39] D. Bogdanov, A. Porter, H. Schreiber, J. Urbano, and S. Oramas, "The acousticbrainz genre dataset: Multi-source, multi-level, multi-label, and large-scale," in *Proc. Int. Symp. Music Information Retrieval*, 2019.

[40] J. Urbano, D. Bogdanov, H. Boyer, E. Gómez Gutiérrez, X. Serra *et al.*, "What is the effect of audio quality on the robustness of MFCCs and chroma features?" in *Proc. Int. Symp. Music Information Retrieval*, 2014, pp. 573–578.

[41] B. McFee, J. W. Kim, M. Cartwright, J. Salamon, R. M. Bittner, and J. P. Bello, "Open-source practices for music signal processing research: Recommendations for transparent, sustainable, and reproducible audio research," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 128–137, 2018.

[42] A. Flexer and T. Grill, "The problem of limited inter-rater agreement in modelling music similarity," *Journal of New Music Research*, vol. 45, no. 3, pp. 239–251, 2016.

[43] K. Seyerlehner, G. Widmer, and P. Knees, "A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems," in *Proc. Int. Workshop Adaptive Multimedia Retrieval*, 2010, pp. 118–131.

[44] J.-J. Aucouturier, "Sounds like teen spirit: Computational insights into the grounding of everyday musical terms," *Language, evolution and the brain*, pp. 35–64, 2009.

[45] M. C. Jones, J. S. Downie, and A. F. Ehmann, "Human similarity judgments: Implications for the design of formal evaluations." in *Proc. Int. Symp. Music Information Retrieval*, 2007, pp. 539–542.

[46] O. Nieto, M. M. Farbood, T. Jehan, and J. P. Bello, "Perceptual analysis of the f-measure for evaluating section boundaries in music," in *Proc. Int. Symp. Music Information Retrieval*, 2014, pp. 265–270.

[47] Y. Ni, M. McVicar, R. Santos-Rodriguez, and T. De Bie, "Understanding effects of subjectivity in measuring chord estimation accuracy," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 12, pp. 2607–2615, 2013.

[48] H. V. Koops, W. B. De Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, pp. 232–252, 2019.

[49] H. V. Koops, W. B. de Haas, J. Bransen, and A. Volk, "Automatic chord label personalization through deep learning of shared harmonic interval profiles," *Neural Computing and Applications*, vol. 32, no. 4, pp. 929–939, 2020.

[50] U. Marchand and G. Peeters, "Scale and shift invariant time/frequency representation using auditory statistics: Application to rhythm description," in *Proc. IEEE Int. Workshop Machine Learning for Signal Processing*, 2016.

[51] D. J. Hand, "Classifier technology and the illusion of progress," *Statistical Science*, vol. 21, no. 1, pp. 1–15, 2006.

[52] G. Peeters, J. Urbano, and G. J. F. Jones, "Notes from the ISMIR 2012 late-breaking session on evaluation in music information retrieval," in *Proc. Int. Symp. Music Information Retrieval*, 2012.