Elsevier Editorial System(tm) for Vision Research

Manuscript Draft

Manuscript Number: VR-15-230R1

Title: Quantitative assessment of intrinsic noise for visually guided behaviour in zebrafish

Article Type: Full Length Article

Keywords: behavioural inconsistency; shoaling; fish cognition; signal detection theory; intraindividual variability

Corresponding Author: Dr. Peter Neri, PhD

Corresponding Author's Institution: Ecole Normale Superieure

First Author: Melissa Spilioti

Order of Authors: Melissa Spilioti; Neil Vargesson; Peter Neri, PhD

Abstract: behavioural inconsistency; shoaling; fish cognition; signal detection theory; intraindividual variability

Cover letter for article VR-15-230

Dear Dr Alais,

thank you for the opportunity to revise this manuscript in response to the constructive comments offered by the Reviewers. We apologize for the unusual delay in revising this submission, but we were keen on performing the additional experiment suggested by Reviewer #2 and, due to logistic difficulties, this was only possible following special arrangements that took a long time to set in place. We have addressed all concerns/comments raised by the Reviewers and we belive that, as a result of this constructive process, our manuscript has greatly improved. In our response letter to the Reviewers, comments by the Reviewers are in boldface. All changes to the manuscript itself have been highlighted in red.

We hope that you find the revised version of the manuscript satisfactory.

Melissa Spilioti
Neil Vargesson
Peter Neri

Response letter for article VR-15-230

1 General comments for Reviewers/Editor

Dear Dr Alais:

Thank you for the opportunity to revise this manuscript in response to the constructive comments offered by the Reviewers. We apologize for the unusual delay in revising this submission, but we were keen on performing the additional experiment suggested by Reviewer #2 and, due to logistic difficulties, this was only possible following special arrangements that took a long time to set in place. We have addressed all concerns/comments raised by the Reviewers and we belive that, as a result of this constructive process, our manuscript has greatly improved. In this response letter, comments by the Reviewers are in boldface. All changes to the manuscript itself have been highlighted in red.

We hope that you find the revised version of the manuscript satisfactory.

Melissa Spilioti
Neil Vargesson
Peter Neri

2 Response to Reviewer #1

Spilioti and colleagues set out to measure intrinsic neural noise in the zebrafish to compare it to similar tests previously done in humans. The shoaling behavior of zebrafish was tested using different contrast patterns depicting shoaling zebrafish. The tested fish were forced to make a choice to swim with two different groups, each representing a contrast group. In general fish preferred the higher contrast group. They demonstrated with a small sample that zebrafish can be used to model behavioral internal noise. This may open the door for future pharmacological experiments seeking to modify this internal noise, though larger studies are likely required to first assess the reliability of this testing paradigm - one of the limitations of this study is that the sample size was quite small.

We agree with the Reviewer that the sample size for this study is relatively small, however we were operating under various constraints (see more detailed response to specific comments below). We have now added a sentence at the end of the manuscript to emphasize this point:

Clearly, far more data than presented here will be necessary to consolidate these tools. Our study represents only a first exploratory step in the direction of identifying whether the proposed tools may be worth pursuing in future research.

We have also performed additional experiments in response to comments by Reviewer #2 using a third cohort of 9 animals (see highlighted text in sections 2.4 and 3.3).

Minor comments: How natural were the movements of the displayed fish? Were the image sprites animated? Tail movement? Etc?

the fish icons were *not* animated, as we now specify within Methods section 2.3:

without any further element of animation (i.e. except for drifting and occasional occlusion by other elements, icons did not undergo any modification). We have demonstrated in previous work that results obtained with actual footage of zebrafish colonies are well-replicated using the artificial stimulus adopted here (Neri 2012).

as explained above, we verified in Neri (2012) that these artificial presentations were as effective as natural footage in driving shoaling behaviour, and more importantly produced the same answers to specific experimental manipulations like stimulus inversion/reverse-playback.

Please provide a statistical analysis section.

Now provided as section 2.6 at the end of Methods.

What is the justification for a 28% failure rate being tolerable? Is there precedence?

We have now expanded the relevant paragraph within Section 3.5 to read:

Across all SNR regimes, the failure rate (∼50%) is substantially higher than observed with human participants; however when restricted to the SNR condition which we identified to be viable on the basis of the above-detailed considerations, the failure rate is in the expected range (2 out of 7 estimates, ∼28%). More specifically, more than ∼10% of human estimates fall outside the viable range even with relatively large trial counts, and failure rate is shown to depend on data mass (Neri 2010a). Because of longer trial duration and behavioural disengagement (see next section), we were able to collect less trials from zebrafish than is typical with humans, which would justify the approximate doubling of observed failures. As for the successful estimates, they are similar to (perhaps slightly higher than) those observed in humans (Burgess & Colborne 1988; Neri 2010a; Diependaele et al. 2012), although more data is required to determine the precise characteristics of this broad agreement.

Why only 7 fish? Fish could be housed individually during the experiment and would allow for greater numbers to be tested? (I recognize that a second cohort of 20 fish were used to some extent, but they could have been housed individually to allow for a more thorough analysis)

As we now explain in Methods section 2.1:

A relevant constraint imposed by ethical guidelines was that fish could not be housed individually for extended periods of time, restricting our ability to identify specific individuals across multiple testing sessions. This guideline is enforced in view of the highly social nature of zebrafish, so as to ensure that they would not be exposed to potentially harming excessive isolation from conspecifics.

We did enquire about the possibility of housing the animals individually, but this was not possible under the unregulated protocol under which we were operating. Individual housing would have required a more extensive ethics application process, which we chose not to engage with at this stage of the project. We therefore opted for anatomical identification of fish housed within the same tank, which we achieved via extensive photograph records of individual animals from different viewpoints. We were able to identify anatomical markers that allowed us to reliably distinguish different animals, however this was only possible for a small number of fish.

We were also operating under additional constraints, for example we could not use all wild type animals in the colony because a large fraction of them was being exploited in other studies and/or breeding. Breeding protocols also imposed restricted hours for access to the facility. At the time at which we carried out these experiments, we were not in a position to collect more data than was done for this study. Since then, both first and senior (last) authors have left the UK, making it nearly prohibitive to collect more data at the present stage. We have nevertheless pushed for an arrangement whereby we were able to perform some additional experiments with a limited (non-ided) cohort in response to a concern raised by Reviewer 2 (please see our response to this Reviewer below), however that was really as far as we could take this study in terms of additional data collection. For future studies, we will need to develop a new set-up with access to a new colony.

Was the vision tested in the fish before experimenting with them? This is commonly done in human testing and could account for some of the variability seen in these fish experiments. Contrast sensitivity and acuity measurements for zebrafish are available (see Tappeiner et l. Frontiers in Zoology 2012, 9:10 doi:10.1186/1742-9994-9-10 and Cameron et al. J Vis Exp. 2013; (80): 50832 doi: 10.3791/50832)

It is uncommon to test visual acuity in fish before experimenting with them (most studies we are aware of in existing literature have not carried out preliminary testing except for those that specifically set out to achieve this aim), so we did not do so. Our data, however, provides strong indication that vision was normal in the animals we tested. We now clarify this issue within Methods (Section 2.4):

The integrity of visual acuity was not explicitly assessed in separate experiments, however the ability of our stimuli to drive all animals under all conditions towards the stimulus with higher mean contrast (data points in Figure 2A fall above the horizontal solid line) is a strong indication that they all possessed neurotypical vision. There were also no visible signs of damage to their eyes, nor swimming behaviour that may indicate (at least on a macroscopic level) impaired visually-guided navigation.

It is possible that inter-individual variability in performance/consistency may correlate with inter-individual differences in acuity as suggested by the Reviewer, however we were not in a position to test this possibility. It remains an interesting avenue for future research.

3 Response to Reviewer #2

This study measured the internal noise of zebrafish using a double-pass paradigm enabling to measure response consistency. The fish were presented with two digital displays containing fish animations on two opposite sides and their preference (i.e., 2afc response) was determined by the side on which they spent the most time. The fish spent more time on the side on which the stimuli (animated fish) were displayed at higher contrast confirming that the fish was responding to the visual stimuli (although performance did not reach 100% at maximal contrast). The authors argued that in one particular condition they succeed at measuring the internal noise, which happen to fall within the humane range. I don't think that the current study convincingly showed that "it is possible to obtain viable estimates of internal noise in this vertebrate species". Further statistical tests (and probably experiments) are required.

As detailed below in response to specific comments by the Reviewer, we have carried out all tests suggested by the Reviewer, including additional experiments/measurements. We hope these important additions are satisfactory to the Reviewer. We emphasize that, as we clarify in the revised submission, our study is not intended as a fully resolved investigation of the relevant issues, but rather as a first step in this direction. Our goal is to identify stimulus parameters and protocol guidelines that could serve as a useful starting point for developing this line of research further. For example, based on our results, future studies would already have a rough idea of what stimulus duration to use, what kind of variability in the estimates to expect, what kind of sample size would therefore be necessary to achieve a certain level of data resolution, what SNR regimes are most likely to yield useful/interpretable outcomes, what hurdles may need to be overcome in designing new protocols (e.g. disengagement with the stimulus as we document it here) and over what timescale, and so forth. In this respect, we believe our study presents valuable material.

1) Excluding the condition in which there was no external noise (SNR=∞, which does not enable to estimate internal noise), three different signal-to-noise ratios (SNR=4, 6 and 12) were tested and gave similar performance levels around 70% correct response. However, only one resulted in a viable estimate of internal noise (SNR=6). This was supported by a Wilcoxon signed-rank test with p<.02. I am not convinced that this is actually significant considering multiple comparisons. Evaluating many different SNR, one will eventually be significant by chance. Proper statistics must be done to show a significant effect considering multiple comparisons.

The p value associated with SNR=6 survives Bonferroni correction for multiple (3x) comparisons (now further clarified by highlighted text within the last paragraph of Section 3.4), but more importantly it should be noted that the test we carried out in Figure 2B is a very stringest test of the viability of our measurements. We now clarify this issue in the Results section 3.4:

The only SNR regime for which percent agreement exceeds the value predicted from percent correct is indicated by red symbols: red data points in Figure 2B fall below the diagonal unity line at p<0.05 on a two-tailed paired Wilcoxon signed rank (WSR) test when Bonferroni-corrected for the 3 multiple comparisons corresponding to the three viable SNR levels (from theory, we do not expect measured percent agreement to be smaller than the stimulus-decoupled prediction, potentially justiying a one-tailed test in this instance, which would strengthen our conclusion). The SNR=∞ condition, indicated by gray symbols, is particularly interesting because it is under this condition that the scenario outlined above would seem most applicable (the two stimuli are perfectly discriminable due to lack of noise); indeed, data points for this condition fall very close to the diagonal unity line. It should be emphasized that the above-detailed test is stringent, because percent agreement values that do not exceed those predicted by the above formula do not imply *that animals were operating in the stimulus-decoupled manner outlined in the previous paragraph: they are consistent with that interpretation, but they also remain consistent with the interpretation based on the standard SDT model. By requiring them to* exceed *the stimulus-decoupled prediction, we are adopting a conservative attitude to exclude for the potential scenario of on-off attentional switching behaviour (see previous paragraph), even though that behaviour may never be applicable to the animals.*

2) The percent agreement necessarily tends to increase with percent correct, so the condition that would be the most likely to observed percent agreement above the predicted percent agreement is when the signal is low and noise is high. Thus, the preferable test conditions would have been with low signal (e.g., μ_1=μ_2=50%) and high noise (e.g., σ=20%), which were unfortunately not performed. Interestingly, the SNR 4 and 6 had the same levels of noise (10%) and different levels of signal (30%-70% vs 20%-80%). Thus, we should expect the method to work better in the lower signal condition (SNR=4). But the results actually showed that the percent agreement was slightly higher than the prediction based on percent correct for SNR=6, but no effect observed at SNR=4. This seems to suggest that the effect observed with SNR=6 may not truly reflect a viable measure. Given that the same level of noise was used in these two conditions, there is no reason for not pooling the data together and statistical test should be performed on the combined data sets.

The Reviewer is correct in the above assertions, however those assertions are based on the assumption that behaviour does conform to SDT. A primary goal of our study was to test precisely this assumption, so that our finding of behaviour conformant with SDT is a contribution in itself. A priori, there is no reason to assume that using a very low signal should return better results: the animal may completely disengage from the stimulus very early in the testing period when signal is too low, and may equally do so when it is too high due to unrewarded contact from the shoal under conditions where there is no competing stimulus and verification of the preferred stimulus is straightforward. It is just not possible to make clear predictions unless one experimentally confirms that SDT applies at least coarsely and at least within some restricted regime. We now clarify these important issues within two newly added paragraphs within Discussion at pages 17-18:

It may seem surprising that stimulus effectiveness did not vary monotonically with SNR: for example, why should the SNR value of 6 work better than values that are both greater and *smaller? Based on SDT considerations, we expect that large SNR values should not be viable, but we also expect that the lower the SNR value, the greater the contribution of external noise, and therefore the more effective the stimulus for internal noise estimation. Indeed, based on SDT considerations alone, a stimulus that only contains noise and no signal should be ideally suited to these experiments. The above considerations are based on the assumption that the behaviour displayed by the animal conforms to our expectations from SDT. There are many alternative scenarios, however. Consider for example the following possibility: that zebrafish may interpret excessive contrast heterogeneity (different icons taking on very different contrast values) as reflecting a non-cohesive shoal where shoal members occupy distant depth planes, and excessive contrast homogeneity (all icons taking the same contrast value) as implausible with unnatural appearance. Under this scenario, stimuli dominated by noise (low SNR) would become less attractive and would drive less shoaling; stimuli dominated by the signal (high SNR) would also drive less shoaling, but for different reasons. The end result in terms of shoaling behaviour as a function of SNR would be difficult to predict and may be non-monotonic.*

We are not suggesting that zebrafish in our experiments were 'interpreting' stimuli as described above, rather we are merely offering one of many example scenarios to illustrate the notion that it would be simplistic to assume that we can predict how the animals will behave as we vary stimulus parameters, because our predictions are based on our own projected model of how the animals 'should' behave. We must first determine the way in which the animals actually *behave; if we then wish to model specific aspects of the observed behaviour, this can only be done within the restricted range for which our model provides a reasonable approximation. In our experiments, for reasons that remain partially unclear at this stage, a stimulus SNR of ∼6 was able to engage the animals with sufficient efficacy to deliver reasonable estimates which cannot be attributed to stimulus-decoupled behaviour (Figure 2B) and that largely conform to SDT.*

To further address the above concerns raised by the Reviewer, we have carried out additional measurements (see point 4 below).

3) The distinction between additive and multiplicative internal noises is not introduced. This is important to more precisely define what is being referred to as "internal noise". I think that in the present context, additive internal noise would correspond to the contrast precision estimate of the samples, whereas multiplicative internal noise would correspond to the noise that is proportional to the combined internal and external additive noises. The double-pass method measures multiplicative noise, not additive noise. The authors need to be more explicit about what is being measured and discuss what intrinsic noise is or is not measured in their paradigm.

We have now added a new paragraph at the end of section 4.3 to define and discuss this specific issue:

When discussing intrinsic noise, the terms 'additive' and 'multiplicative' are often adopted to label different types of internal variability. This terminology can be misleading, however, because the same source of behavioural variability may be incorporated as additive or multiplicative by different models, so that model architecture becomes critical for drawing the additive/multiplicative distinction. In the standard SDT model adopted here, noise consists of a Gaussian fluctuation added to the decisional variable. In this sense, it is late additive ('late' refers to the stage at which it is added, this being the last stage before producing a behavioural response). Its unit, however, is the standard deviation of the distribution taken by the decisional variable as a result of external *stimulus noise (this is also how d′ is defined): its intensity is therefore defined as a multiple of external noise, potentially generating confusion. For example, if external noise is varied and the estimated value of internal noise remains unchanged (as is typically found (Neri 2010b)), this means that the intensity of internal noise has actually changed and it has done so in a manner that scales proportionally with external noise by the same constant value. More importantly, because internal noise as defined and estimated here potentially encompasses multiple sources of internal variability (see above), it is not possible to know with certainty whether its physiological origin is additive or multiplicative in nature. Our choice of model is motivated by extensive literature justifying its general applicability to human vision (Burgess & Colborne 1988; Neri 2010a; Neri 2013).*

4) Furthermore, quantifying multiplicative internal noise proportionally to the external (additive) noise implies assuming that performance was driven by external noise, not internal additive noise. Otherwise, the external noise would have negligible impact and it would be impossible to quantify the multiplicative internal noise relative to external noise. A simple way to test whether additive internal or external noise drives performance is to measure performance with and without the external noise with the same signal level. Unfortunately, the condition in which there was no noise had a different signal strength. Nonetheless, two conditions (SNR=6 and 12) had the same signal strength but different noise levels (5 and 10%). Significantly different performances would show that the noise noticeably affected performance. Authors need to show that the external noise had an impact on performance since the important measure is quantified relative to the impact of this noise.

We sincerely thank the Reviewer for making this point, and for prompting us to perform additional experiments to directly compare noisy versus noiseless conditions that only differed in the presence/absence of noise, i.e. with matched mean-contrast separation. It was logistically difficult to perform these additional experiments due to the senior author (PN) leaving the University of Aberdeen and the lead author (MS) finishing her studies, which explains the unusual delay in revising this article. We were able to come up with an arrangement whereby we could test an additional cohort of 9 wild-type animals that we could not individually label, but for which we ran a direct comparison between configuration SNR=6 and the same configuration without external noise (this additional cohort is now described in section 2.4). As detailed in the manuscript (see below), the results of these additional experiments unambiguously confirmed that external noise did impact behaviour, as now detailed in section 3.3:

The framework outlined above, whereby internal noise is expressed as a multiple of the variability generated by externally applied noise, rests on the assumption that the external noise source is having a measurable impact on behaviour: if not, all variability is internally generated, and it cannot be defined as a multiple of a quantity that is 0. We return to this point in relation to the notion of stimulus-decoupled behaviour (see below). To directly gauge the validity of this assumption, we measured preference for the higher-mean-contrast stimulus on a sample of 9 animals presented with two different stimulus configurations having equal mean-contrast difference between the two competing stimuli, but either no external noise in one configuration, and external noise corresponding to SNR=6 in the other configuration. More specifically, the SNR=6 configuration was identical to the one detailed previously and pictured in Figure 1B, while the configuration without external noise contained no contrast variability from fish to fish within a given movie (similar to Figure 1D) but a mean-contrast separation that matched the mean separation used for the SNR=6 stimulus. If the externally applied noise source (in the form of contrast changes from fish to fish in the stimulus) does not impact behaviour, we expect comparable preference for the higher-mean-contrast stimulus under these two different configurations; if, on the other hand, the application of external noise did impact behaviour, we expect reduced preference in the presence of external noise, due to the lower discriminability (SNR=6 as opposed to SNR=∞) associated with the presence of external noise. Our results unequivocally confirmed the latter expectation: preference was in the range (minimum/median/maximum) of 0.45/0.65/0.75 across the 9 animals tested for the SNR=6 condition, while it measured 0.7/0.75/0.9 for the condition without external noise. The difference between the two conditions was statistically significant at $p < 0.0005$ *(unpaired two-tailed Wilcoxon rank sum test), clearly indicating that external noise as designed and applied in our protocols did impact visually-guided behaviour of the test animal, and in turn supporting definition of internal noise within the framework outlined in the previous paragraph.*

5) It is mentioned a few times that the internal noise is measured in units of external noise and therefore, cannot be quantified when the external noise was 0 as when the SNR=∞, but the values of internal noise when the external noise was 0 are represented in Figure 3. How could that be? How was the internal noise calculated in this condition and what does it represent? Maybe I'm missing something$\cdots$

The Reviewer is correct, and the SNR=∞ condition was merely included as a 'sanity check'. We now clarify this point further at the end of the first paragraph of section 3.4:

In other words, our goal was to verify that our analysis tools would be able to exclude this condition as viable even though the associated empirical measurements may still be fed to the estimation algorithm and generate outputs (as discussed later in the article, we find indeed that the resulting internal noise estimates are well within the failure range and that none of the measured percent agreement values for this condition exceed those expected of stimulus-decoupled behaviour).

6) What was the stimulus used in the disengagement experiment? (SNR=6 I suppose). Please specify.

now specified at the beginning of the second paragraph of section 3.7.

7) Results section. Many subsection of the results section do not describe results but methodology or general concepts. This should be substantially revised.

We attempted to move some sections out of the Results section, however we chose to retain a lot of the original structure because, in our experience, many readers skip the Methods section upon initially approaching the paper, only referring to it in the event of clarifying specific details. Our intention was therefore to offer enough description of the general methodology underlying our results in a manner that enabled a direct link between the results and the adopted methods, so that potentially the Results section would stand on its own. We understand that this stylistic choice is not favoured by all readers, and we apologize to this Reviewer if he/she feels that it is inadequate, but we eventually decided to retain it because completely splitting method description and results for a study like this one, where different stimuli/conditions are used, required prospective readers to constantly switch between Results and Methods sections while keeping track of what results go with what stimuli/protocols, a situation we wished to avoid.

8) Section 3.2. I don't think that the relationship between sensitivity and consistency should be described as a result. There is necessarily a link between performance and consistency (performance of 100% implies consistency of 100%). This is well known and should be introduced with the model in the introduction, not presented as a result.

As we have explained in response to comment 2 above by the Reviewer, a primary goal of our study was to gauge the applicability of SDT for visually guided behaviour in the zebrafish. For this system, the result is not well-known unless one assumes that SDT applies within that context, which we set out to verify. It is certainly the case that, as the Reviewer points out, perfect performance *must* correspond to perfect agreement, but for intermediate values there is a whole set of trends that may be expected depending on what the animal is doing; our observation as a first step in interpreting the plot was that, at a coarse level, the observed trend conforms with SDT predictions. We have now added a clarification in this respect that specifically addresses the comment above:

The latter trend is expected from signal detection theory (SDT) (Burgess Colborne 1988), however this expectation does not trivialize the empirically observed trend: one goal of this study is to establish whether visually-guided behaviour in the zebrafish can at all *be approximated by SDT in the first place. Our observation of compatible characteristics between measured behaviour and SDT therefore provides added knowledge beyond what is available from current literature: there are no prior measurements of response agreement in zebrafish; without measuring this quantity directly, it remains conceivable that a different trend may have been observed (see further discussion of this issue below in relation to the interpretability of specific behavioural patterns and their relationship to stimulus parameters)*

this addition is within the last paragraph of section 3.2.

Quantitative assessment of intrinsic noise for visually guided behaviour in zebrafish

Melissa Spilioti

Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK

Neil Vargesson

Institute of Medical Sciences, University of Aberdeen, Aberdeen, UK

Peter Neri

Laboratoire des Systèmes Perceptifs (CNRS UMR 8248) and Département d'études cognitives, Ecole Normale Supérieure, PSL Research University, Paris, France

Corresponding author: Peter Neri, neri.peter@gmail.com

Abstract: All sensory devices, whether biological or artificial, carry appreciable amounts of intrinsic noise. When these internally generated perturbations are sufficiently large, the behaviour of the system is not solely driven by the external stimulus but also by its own spontaenous variability. Behavioural internal noise can be quantified, provided it is expressed in relative units of the noise source externally applied by the stimulus. In humans performing sensory tasks at near threshold performance, the size of internal noise is roughly equivalent to the size of the response fluctuations induced by the external noise source. It is not known how the human estimate compares with other animals, because behavioural internal noise has never been measured in other species. We have adapted the methodology used with humans to the zebrafish, a small teleost that displays robust visually-guided behaviour. Our measurements demonstrate that, under some conditions, it is possible to obtain viable estimates of internal noise in this vertebrate species; the estimates generally fall within the human range, suggesting that the properties of internal noise may reflect general constraints on stimulus-response coupling that apply across animal systems with substantially different characteristics.

Keywords: behavioural inconsistency — shoaling — fish cognition — signal detection theory — intraindividual variability

1 INTRODUCTION

Biological systems do not behave deterministically: when presented with two identical instances of an external event, they may react differently depending on their internal state at the time of stimulation (Green 1964; Highcock & Carter 2014). This observation applies without exception to conditions where a stimulus signal is corrupted by an external noise source, and a human participant is asked to detect the presence of the signal: identical instances of signal and noise will result in different reports on the part of the human participant on about 3 out of 4 stimulus replications (Burgess & Colborne 1988; Neri 2010a).

It is possible to measure this departure from deterministic behaviour and quantify the amount of internal perturbation, but this can only be done in a relative sense. Because behaviour is driven by the internal representation of the stimulus, internal noise can only be defined with relation to this internal representation, which lacks absolute units. In the dominant framework for the quantification of animal behaviour, termed signal detection theory (SDT), this issue is addressed by rescaling all perceptual quantities (e.g. sensitivity) as a function of the variability induced upon them by variations within the external stimulus (Green & Swets 1966). The same approach can be applied to internal noise (Burgess & Colborne 1988; Neri 2010a), thus enabling estimates of this phenomenon that are not only quantitative, but in principle directly comparable across different species provided sensory behaviour for the species in question can be adequately modelled using the principles of SDT.

In light of the above-stated potential for comparative studies of a fundamental property of animal behaviour such as internal noise, it may seem surprising that this phenomenon has so far been quantified only in humans. To our knowledge, there have been no comparable measurements in other species, making it difficult to interpret the human measurements on a broader scale that takes into account their comparative significance. Intra-individual variability (IIV), a quantity commonly used to study related phenomena (MacDonald *et al.* 2006), lacks an established theoretical framework (Biro & Adriaenssens 2013); its potential for comparative judgements is therefore compromised by the unavailability of a common metric space across different species. The goal of our experiments was to rectify these limitations and allow for direct comparison of intrinsic behavioural noise between humans and a small vertebrate, the zebrafish, that has proven a useful animal model for genetic manipulations relating to a range of human pathological conditions (Norton & Bally-Cuif 2010), some of which (ADHD in particular) are believed to stem from abnormalities associated with internal noise (Gilden & Hancock 2007; Simmons *et al.* 2009; Perry *et al.* 2010; Dinstein *et al.* 2012; Kofler *et al.* 2013).

2 METHODS

2.1 Animals and test apparatus

Except for the visual stimuli, which were specifically designed for this study (see next section), all other procedures were identical to those described in previous work (Neri 2012) and will only be summarized here. We used wild-type zebrafish bred and maintained by trained staff in a dedicated facility (Institute of Medical Sciences, Aberdeen, United Kingdom; see also Vargesson 2007; Theraponts & Vargesson 2010 for details relating to husbandry). Outside testing, fish were kept inside a 10-litre storage tank (average density two fish per litre) attached to a recirculated system (Aquatic Habitats, Apopka, FL, U.S.A.) at 27°C on a 14:10 h light:dark photoperiod and never exposed to heterospecifics. They were fed brine shrimp twice a day (at 09:30 and 16:30). During testing, one fish was transferred from the facility to a test tank measuring 25×13 cm and 11 cm high. The two furthest sides of the test tank were placed against two identical LCD monitors driven by one computer allowing independent control over the images displayed to the two sides. A webcam located above the test tank acquired images at 4 Hz and stored them on the hard drive for automated offline analysis. After testing, fish were returned to the breeding stock. Ethical approval for all the research reported in this study was obtained from the University of Aberdeen Ethical Review Committee. The work, which was in accordance with the Code of Ethics of the World Medical Association (Declaration of Helsinki), was deemed as nonregulated by the Home Office Inspector; however, input was received from the Home Office Inspector and the Named Veterinary Surgeon and the care of all fish was under the remit of the Animals (Scientific Procedures) Act 1986. No animal licence was required because the behavioural procedures used here were non-invasive, in accordance with natural behaviour patterns, and only involved wild-type animals. A relevant constraint imposed by ethical guidelines was that fish could not be housed individually for extended periods of time, restricting our ability to identify specific individuals across multiple testing sessions. This guideline is enforced in view of the highly social nature of zebrafish, so as to ensure that they would not be exposed to potentially harming excessive isolation from conspecifics.

2.2 Automated tracking of animal position

We wrote software specifically tailored to the images collected during the experiments; the algorithm was therefore robust and efficient in the absence of any human intervention. Readers are referred to (Neri 2012) for details. Briefly here, the software implemented motion detection via thresholded

subtraction methods (McIvor 2000) and applied cluster analysis to identify the test animal. The location of the cluster centroid between automatically detected end-points for the tank was used as position marker (see red/blue dots in Figure 1E). To determine whether the test animal preferred one or the other side of the tank on a specific trial, we simply averaged all position values over the duration of that trial (see red/blue lines in Figure 1E); preference was assigned to the side of the tank closest to this average value. We also explored other methods for assigning preference, for example the % time spent on either side of the tank, but this had no appreciable impact on our results. Furthermore, we were not able to expose any systematic relationship between the specific value of mean (or median) shift displayed by the animal on individual trials and the mean contrast difference of the stimuli presented on those same trials. In other words, although the mean contrast difference systematically modulated the preference as assessed via probability of binary choice, it did not appear to modulate the mean shift on a given trial, or at least not within the resolution of our measurements.

2.3 Visual stimuli and presentation protocol

All stimuli were generated by adding the same small icon of a zebrafish to a grey background. Ten individual icons were initially placed within the image at random spatial locations and made to drift horizontally at a constant speed of 6.5 cm/s without any further element of animation (i.e. except for drifting and occasional occlusion by other elements, icons did not undergo any modification). We have demonstrated in previous work that results obtained with actual footage of zebrafish colonies are reliably replicated using the artificial stimulus adopted here (Neri 2012). Half the icons moved to the left and half to the right. When two icons overlapped within the image, the icon added more recently was painted over the other icon. All movies lasted 16 s and were generated using a cyclical structure: the end of the movie matched the beginning of the movie, so that the movie could be played smoothly for multiple repetitions without glitches. For a given movie, the contrast of each icon was randomly drawn from a Gaussian distribution with mean μ_j and standard deviation σ, where j is 1 for the movie with higher mean contrast and 2 for the movie with lower mean contrast (i.e. $\mu_1 > \mu_2$). Both high and low mean-contrast movies were presented during each trial on opposite sides of the tank; which side contained the high contrast movie was randomly determined. On a given test lasting $\sim$14 minutes, the animal was presented with 1 block of 20 trials. Each trial lasted 30 seconds, and trials were separated by a 10-second gap during which both monitors displayed blank screens. Each block was associated with a specific parameterization (μ_1, μ_2 and σ values) of the contrast distributions defining the two stimuli; each parameterization corresponds to

a different signal-to-noise ratio (SNR) $(\mu_1 - \mu_2)/\sigma$. We tested 4 different SNR values: 4 defined by μ_1=70%, μ_2=30% and σ=10% contrast (Figure 1A); 6 defined by μ_1=80%, μ_2=20% and σ=10% contrast (Figure 1B); 12 defined by μ_1=80%, μ_2=20% and σ=5% contrast (Figure 1C); ∞ defined by μ_1=100%, μ_2=0% and σ=0% contrast (Figure 1D). Each block was divided into two 'passes': the 1st pass from trial #1 to trial #10, the 2nd pass from trial #11 to trial #20. The stimulus samples presented during the 1st pass were independently generated: on trial #1, the stimulus on the right side of the tank may contain 10 fish with contrast values randomly drawn from the distribution with higher mean μ_1, while the stimulus on the left side would then contain 10 fish with contrast values randomly drawn from the distribution with lower mean μ_2 (see icons on top row of Figure 1E); on trial #2, the stimulus on the right may still draw from the contrast distribution with higher mean (see icons on second row of Figure 1E), but it would be a different random sample, and so would be the stimulus on the other side; on trial #3, the stimulus on the right side may now draw from the contrast distribution with lower mean (see icons on third row of Figure 1E), and so on. The 2nd pass was an exact replication of the 1st pass: the same stimulus samples were presented on the same side of the tank as during the 1st pass.

2.4 Number of test animals and data mass

We tested three different cohorts. The first cohort consisted of 7 animals (age range 1.5-2 years old) which we could identify individually based on specific morphological features (e.g. irregularities of their stripe pattern, body asymmetries); we were restricted in our ability to test a large number of individually identifiable animals due to a combination of ethical guidelines (see above) and breeding requirements within the facility. The integrity of visual acuity was not explicitly assessed in separate experiments, however the ability of our stimuli to drive all animals under all conditions towards the stimulus with higher mean contrast (data points in Figure 2A fall above the horizontal solid line) is a strong indication that they all possessed neurotypical vision. There were also no visible signs of damage to their eyes, nor swimming behaviour that may indicate (at least on a macroscopic level) impaired visually-guided navigation. For stimulus SNR=4, we collected 2 blocks from each of 5 animals and 1 block from each of the remaining 2 animals (total of 12 blocks); for SNR=6, we collected 5 blocks from each of 6 animals and 3 blocks from the remaining animal (total of 33 blocks); for SNR=12, we collected 1 block from each animal (total of 7 blocks); for SNR=∞, we collected 2 blocks from each animal (total of 14 blocks). We allocated more data collection to condition SNR=6 because piloting indicated that this condition returned more robust estimates from individual blocks than the remaining three conditions. This preliminary indication was confirmed by

further analysis, as demonstrated in Figures 2-3. Notice that the estimates reported in those figures were obtained by first computing an estimate from each block and then averaging across blocks, not by first collating trials across different blocks. The second cohort consisted of 20 animals (similar age to the first cohort) which we could not identify individually. We collected 1 block from each animal at SNR=6. The results we obtained from this second cohort closely matched those obtained from the first cohort (compare Figure 4B with A; see also open circle in Figure 3). The third cohort consisted of 9 animals (similar age to the other two cohorts) which we could not identify individually. We collected 1 block from each animal at SNR=6, and 1 additional block for a configuration with equal mean-contrast separation between the two stimuli, but no external noise. These two configurations were specifically selected to differ only in the presence/absence of external noise, so that the impact of external noise could be gauged directly.

2.5 Estimation of internal noise

Our methodology relies on the established signal detection theory (SDT) model (Green & Swets 1966). The SDT model is defined within the space of the 'internal response': the response of the system to the input stimulus, regardless of the front-end process that maps the stimulus onto a response. This process may consist of the human visual system or the zebrafish visual system; the details are not relevant because the SDT formulation bypasses this stage. For our 2AFC task, we assume that the internal response before the addition of internal noise follows a normal distribution for the nontarget low-mean-contrast stimulus and a normal distribution with mean d'_{in} for the target high-mean-contrast stimulus. Each response is added to a Gaussian noise source with SD σ_N; only this noise source differs for repeated presentations of the same stimuli on the two passes, and represents internal noise. On each trial, the model selects the stimulus associated with the largest response. d'_{in} and σ_N are not directly measurable: they are model parameters. However, different d'_{in} and σ_N values correspond to different values of two directly measurable quantities: percent correct and percent agreement (Burgess & Colborne 1988). The % of correct responses is the % of trials on which the animal showed preference for the side of the tank displaying the stimulus defined by the higher contrast mean. Agreement is the % of paired trials associated with the same preference on the two passes: preference on the first trial of the 1^{st} pass (trial #1 within the block) is matched against preference on the first trial of the 2^{nd} pass (trial #11 within the block), preference on the second trial of the 1^{st} pass (trial #2 within the block) is matched against preference on the second trial of the 2^{nd} pass (trial #12 within the block), and so on. The % of matches is percent agreement. We then selected the specific values for d'_{in} and σ_N that minimized the mean-square error between

the predicted and the observed values for percent correct and percent agreement (Neri 2010a). The orange lines in Figure 2A define pairings of percent-correct/percent-agreement values corresponding to different d'_{in} values (as one moves along the line) for a fixed σ_N value (indicated below each line).

2.6 Statistical analysis

With the exception of p values from correlation tests, obtained via the t-statistic, all other p values come from two-tailed non-parametric Wilcoxon tests (paired when involving comparisons between two samples, except for one test relating to the third cohort where the comparison between the two samples could not be paired due to the lack of individually identified data, and it was therefore unpaired). Bonferroni correction for multiple comparisons is adopted when applicable.

3 RESULTS

3.1 Stimulus parameterization

Zebrafish exhibit a spontaneous form of visually-guided behaviour termed 'shoaling', whereby exposure to real or simulated images of conspecifics results in an innate tendency towards aggregation (Miller & Gerlai 2011). This phenomenon can be exploited to support experimental conditions that mirror classic two alternative forced choice (2AFC) protocols from visual psychophysics (Orger *et al.* 2000; Engeszer *et al.* 2004; Neri 2012): the animal is presented with two different visual stimuli on opposite sides of the tank, each containing a manipulated movie depicting conspecifics, while its position is tracked to monitor its tendency to spend more time on one side of the tank as opposed to the other. Preference can be coded as a binary variable: 1 if the animal spends more time on the side of the tank associated with stimulus number 1; 2 if it spends more time on the other side. Under these conditions the fish is essentially performing a 2AFC task, enabling deployment of a large body of established techniques from visual psychophysics (Green & Swets 1966; Burgess & Colborne 1988; Neri 2010a).

Our stimulus consisted of a synthetic zebrafish shoal (Saverino & Gerlai 2008; Neri 2012). We used the same image for all 10 members of the synthetic shoal, but varied the contrast of each member independently. For a given shoal sample, the 10 contrast values assigned to the different members were sampled from a Gaussian distribution. We manipulated stimulus discriminability by separately specifying mean and standard deviation of the distributions underlying the two stimulus classes presented to the fish. Stimulus discriminability is defined as the difference between the two means divided by their common standard deviation (see Methods): the signal-to-noise ratio (SNR)

(Green & Swets 1966). We tested 4 different stimulus parameters associated with different stimulus SNR values (4, 6, 12 and ∞; see Figure 1). As detailed below, we found that only one of these 4 conditions (SNR=6) supported behavioural regimes that allowed for adequate measurements of internal noise in the zebrafish.

3.2 Relationship between sensitivity and consistency

On each test (lasting $\sim$14 minutes), we presented 10 different pairs of samples for stimulus 1 and 2. Each sample pair was presented twice on two different trials. As shown in Figure 1, when the animal was presented with a repeated stimulus pair, it did not always display the same preference on the two presentations (compare red and blue trajectories in Figure 1E). The percentage of trials on which preference was consistent (i.e. the same on both presentations) was not at chance (50%), but was not perfect either (i.e. it never reached 100%). Figure 2A plots this quantity on the x axis for different animals (identified by different symbols) and different stimulus SNR's (indicated by different colours; see also Figure 1A-D). The y axis plots the corresponding percentage of trials on which the animal displayed preference for the stimulus with higher mean contrast. In keeping with established literature, these two quantities may also be termed consistency and sensitivity respectively (Burgess & Colborne 1988).

Before proceeding to a quantitative evaluation of the data in Figure 2A, we notice a few qualitative features of the manner in which data points scatter across the plot. First, all data points bar one fall above the horizontal black line corresponding to unbiased behaviour (0.5), demonstrating that zebrafish displayed preference toward the higher-contrast stimulus. Furthermore, the average y position of the different datasets corresponding to different SNR values (different colours) shifts upwards with increasing SNR (see arrows pointing towards left y axis), demonstrating that our visual stimuli were able to drive behaviour in a lawful manner. Third, most data points fall to the right of the vertical black line corresponding to chance agreement between repeated presentations, demonstrating that zebrafish showed a measurable degree of consistent behaviour. Finally, values on the two axes covary positively: larger sensitivity values are associated with larger consistency values.

The latter trend is expected from signal detection theory (SDT) (Burgess & Colborne 1988), however this expectation does not trivialize the empirically observed trend: one goal of this study is to establish whether visually-guided behaviour in the zebrafish can *at all* be approximated by SDT in the first place. Our observation of compatible characteristics between measured behaviour and SDT therefore provides added knowledge beyond what is available from current literature: there are no prior measurements of response agreement in zebrafish; without measuring this quantity directly,

it remains conceivable that a different trend may have been observed (see further discussion of this issue below in relation to the interpretability of specific behavioural patterns and their relationship to stimulus parameters). The orange lines plot predicted relationships between percent correct and percent agreement for different degrees of intrinsic noise associated with a system that behaves according to a minimal SDT model. These predictions demonstrate that consistency and sensitivity are indeed expected to covary positively, further corroborating the notion that our dataset presents meaningful structure and that this structure can be modelled and understood using the established tools of statistical decision theory (Green & Swets 1966).

3.3 Zebrafish as SDT operators

The above observations suggest that, at least to a coarse extent, visually-guided behaviour in the zebrafish may be approximated by the general framework associated with SDT. Within the context of SDT, internal noise is measured in units of the perceptual fluctuations induced by the external noise source (Burgess & Colborne 1988; Neri 2010a). To understand this concept, imagine that each stimulus in Figure 1E is associated with a perceptual response of a given intensity within the sensory machinery of the animal (Diependaele *et al.* 2012). Because this response is defined in perceptual space, we cannot express it in absolute units: perceptual space has no units like spikes per second or BOLD signal intensity. This issue is easily addressed by redefining all quantities as multiples of (i.e. in units of) the variability associated with the perceptual response (i.e. its standard deviation). To provide a relevant example, the discriminability between two stimuli, i.e. the difference in perceptual response to those two stimuli (which underlies behavioural sensitivity) is divided by the variability of the two responses to obtain d' (Green & Swets 1966).

Response variability comes from two sources: the variability introduced by the external stimulus which contains noise in the form of contrast fluctuations (Figure 1A-C), and the additional variability introduced by the intrinsic noisiness of the animal (inconsistency; Green 1964; Burgess & Colborne 1988; Diependaele *et al.* 2012). Because variability is used as unit of measurement in perceptual space, it does not make sense to speak of variability itself in those units; it is only the relative intensity of the two sources that we can meaningfully quantify and estimate: we can say, for example, that total variability is due to external noise for 25% of its intensity, and to internal noise for the remaning 75%. This would mean that internal noise is $3\times$ the external noise souce. In humans, the intensity of internal noise falls between $1/2$ and 2, i.e. it may be as low as half the external noise source and as large as twice its value (Neri 2010a). The latter case is represented by the darker orange line in Figure 2A. In the next section, we examine how the different SNR datasets relate to this upper

boundary on the human range.

The framework outlined above, whereby internal noise is expressed as a multiple of the variability generated by externally applied noise, rests on the assumption that the external noise source is having a measurable impact on behaviour: if not, all variability is internally generated, and it cannot be defined as a multiple of a quantity that is 0. We return to this point in relation to the notion of stimulus-decoupled behaviour (see below). To directly gauge the validity of this assumption, we measured preference for the higher-mean-contrast stimulus on a sample of 9 animals presented with two different stimulus configurations having equal mean-contrast difference between the two competing stimuli, but either no external noise in one configuration, and external noise corresponding to SNR=6 in the other configuration. More specifically, the SNR=6 configuration was identical to the one detailed previously and pictured in Figure 1B, while the configuration without external noise contained no contrast variability from fish to fish within a given movie (similar to Figure 1D) but a mean-contrast separation that matched the mean separation used for the SNR=6 stimulus. If the externally applied noise source (in the form of contrast changes from fish to fish in the stimulus) does not impact behaviour, we expect comparable preference for the higher-mean-contrast stimulus under these two different configurations; if, on the other hand, the application of external noise did impact behaviour, we expect reduced preference in the presence of external noise, due to the lower discriminability (SNR=6 as opposed to SNR=∞) associated with the presence of external noise. Our results unequivocally confirmed the latter expectation: preference was in the range (minimum/median/maximum) of 0.45/0.65/0.75 across the 9 animals tested for the SNR=6 condition, while it measured 0.7/0.75/0.9 for the condition without external noise. The difference between the two conditions was statistically significant at $p < 0.0005$ (unpaired two-tailed Wilcoxon rank sum test), clearly indicating that external noise as designed and applied in our protocols did impact visually-guided behaviour of the test animal, and in turn supporting definition of internal noise within the framework outlined in the previous paragraph.

3.4 Internal noise estimation is only viable within a restricted SNR range

Of the four different SNR regimes we tested, only that associated with the red dataset in Figure 2A approaches the upper boundary of the human range (and sometimes falls below it). The dataset for the SNR value immediately below (black symbols) occasionally falls within this range, but some estimates (black circle and downward triangle) are associated with percent agreement measurements below chance (left of vertical black line) and are therefore incompatible with the SDT model (Green & Swets 1966). The higher SNR regimes (blue and gray symbols) return datasets that scatter in the

region of infinite values for internal noise (thick light-orange line in Figure 2A) and are therefore also not viable for the purpose of sensible estimation. This is expected for stimuli containing no contrast fluctuations (SNR=∞, gray dataset) because the external noise source has 0 standard deviation, making it impossible to express internal noise as relative to external noise. Our motivation for testing this SNR condition was for the resulting measurements to serve as a sanity check that our methods and analyses would integrate meaningfully across the board, even under limit conditions for the relevant parameters. In other words, our goal was to verify that our analysis tools would be able to exclude this condition as viable even though the associated empirical measurements may still be fed to the estimation algorithm and generate outputs (as discussed later in the article, we find indeed that the resulting internal noise estimates are well within the failure range and that none of the measured percent agreement values for this condition exceed those expected of stimulus-decoupled behaviour).

A potentially puzzling feature of Figure 2A is that the model prediction associated with unreasonably large (nearly infinite) internal noise intensity (thick light-orange line) produces percent agreement values exceeding chance; this may seem nonsensical, because consistency should be near chance when internal noise is huge. The regime we are considering lies near the limit case of infinite internal noise, when indeed both consistency and sensitivity should be at chance. In the vicinity of the limit case, it is instructive to consider the problem from a slightly different perspective (see below).

Imagine the system responds correctly on $x\%$ of trials, but its behaviour bears no relationship to the discriminability of individual stimulus samples: the system merely responds correctly on a randomly chosen subset of trials. A possible scenario that would generate this type of behaviour is one where the animal ignores the presented stimuli on some trials, and thus responds randomly on those trials, but pays great attention to the stimuli presented on the remaining trials, and thus discriminates those with near-perfect accuracy. Under these conditions (violating the basic assumptions of SDT), a specific stimulus pair is no more likely to cause the same behaviour on its repeated presentation than is expected on an unrelated trial, which would correspond to infinite internal noise; percent agreement, however, will not be at chance: if p is the probability that any trial is associated with a correct response, the probability that both repetitions will yield the same response (whether correct or incorrect) is $p^2 + (1-p)^2$. In order for a percent agreement value to reflect true behavioural consistency, rather than potentially being the byproduct of a higher-than-chance percent correct value, it is therefore necessary that it exceeds the value returned by this expression. The corresponding viable region is indicated by green shading in Figure 2B, where the outcome of the above-detailed expression (computed by simply replacing p with the measured

percent correct values) is plotted on the y axis versus the empirically measured percent agreement values (replotted from the x axis in Figure 2A).

The only SNR regime for which percent agreement exceeds the value predicted from percent correct is indicated by red symbols: red data points in Figure 2B fall below the diagonal unity line at p=0.016 on a two-tailed paired Wilcoxon signed rank (WSR) test; when Bonferroni-corrected for the 3 multiple comparisons corresponding to the three viable SNR levels, this remains significant at $p<0.05$ (from theory, we do not expect measured percent agreement to be smaller than the stimulus-decoupled prediction, potentially justiying a one-tailed test in this instance, which would strengthen our conclusion). The SNR=∞ condition, indicated by gray symbols, is particularly interesting because it is under this condition that the scenario outlined above would seem most applicable (the two stimuli are perfectly discriminable due to lack of noise); indeed, data points for this condition fall very close to the diagonal unity line. It should be emphasized that the above-detailed test is stringent, because percent agreement values that do not exceed those predicted by the above formula do not *imply* that animals were operating in the stimulus-decoupled manner outlined in the previous paragraph: they are consistent with that interpretation, but they also remain consistent with the interpretation based on the standard SDT model. By requiring them to *exceed* the stimulus-decoupled prediction, we are adopting a conservative attitude to exclude for the potential scenario of on-off attentional switching behaviour (see previous paragraph), even though that behaviour may never be applicable to the animals.

3.5 Explicit estimates of internal noise

As we have explained with relation to Figure 2A, different parameterizations of the SDT model are associated with different predictions for the relationship between percent agreement and percent correct values (Burgess & Colborne 1988; a representative sample of four different predictions is indicated by orange traces). Based on the experimentally observed values, we can derive estimates for the best-fitting parameters within the underlying SDT model (Neri 2010a; Diependaele *et al.* 2012; see Methods). This model is defined by two parameters: stimulus discriminability and internal noise (both in units of external noise standard deviation). They are plotted in Figure 3 on x and y axes respectively.

In approaching this dataset, it seems useful to rely on related measurements in human participants for general guidance. Previous work with large-scale datasets has demonstrated that $\sim$90% of internal noise estimates fall between 1/5 and 5 (margins indicated by orange horizontal dashed lines in Figure 3); estimates outside this range are most reasonably regarded as failures of the adopted

methodology and should be excluded from further consideration (Neri 2010a). The representative range for human sensory processing is between 0.6 and 2, indicated by green shading in Figure 3.

In line with the results detailed earlier, only the SNR regime associated with the red dataset returned a majority of estimates within the acceptable range; interestingly, the estimates that did fall within this region also clustered within the representative human range (green shading in Figure 3). Across the entire dataset, internal noise estimates were distributed bimodally (see histogram to the right) with two populations on opposite sides of the upper boundary for the viable region (value of 5 on y axis, indicated by top orange horizontal dashed line). We interpret this bimodality as reflecting well-segregated successes/failures of our methodology: our protocols either succeed (estimates < 5) or fail (estimates > 5).

Across all SNR regimes, the failure rate ($\sim$50%) is substantially higher than observed with human participants; however when restricted to the SNR condition which we identified to be viable on the basis of the above-detailed considerations, the failure rate is in the expected range (2 out of 7 estimates, $\sim$28%). More specifically, more than $\sim$10% of human estimates fall outside the viable range even with relatively large trial counts, and failure rate is shown to depend on data mass (Neri 2010a). Because of longer trial duration and behavioural disengagement (see next section), we were able to collect less trials from zebrafish than is typical with humans, which would justify the approximate doubling of observed failures. As for the successful estimates, they are similar to (perhaps slightly higher than) those observed in humans (Burgess & Colborne 1988; Neri 2010a; Diependaele *et al.* 2012), although more data is required to determine the precise characteristics of this broad agreement.

3.6 Animals disengage with the stimulus over time

We noticed a consistent trend whereby preference on the part of the test animal was more effectively driven by our stimulus at the beginning of each experiment and gradually decreased over time. Figure 4A plots the percentage of trials on which the animal shoals towards the high-contrast stimulus separately for each of four different epochs within each block: the first 5 trials of the 20 trials that contributed to a given block, the second 5 trials (6-10), and so on. Because we could identify specific individuals, we were in a position to combine data from different experimental sessions and plot the results separately for different animals. All 7 fish present a negative (or near 0) trend of performance with trial progression (a two-tailed Wilcoxon test for the 7 correlation values being different from 0 returns $p<0.02$; all linear fits in Figure 4A present a negative slope). For some individuals (upper triangles in Figure 4A) the animal was well above chance ($\sim$0.8) at the beginning of the test, and

reached chance performance by the end of the block.

We wished to confirm this trend in a larger cohort of different individuals. We therefore tested an additional 20 fish, none from the previous population, for one block each (see Methods). Because we were not in a position to identify specific individuals within this cohort, we could not collate data across sessions and we therefore collected only one block (20 trials) per individual. Due to the more limited amount of data available for each individual, it was not possible to perform the analysis separately for each individual; we therefore plot the aggregate result (across individuals) in Figure 4B. The advantage with respect to the plot in Figure 4A is that, because we are averaging across individuals and there was a larger number of them, we can resolve the trend with 1-trial resolution. The negative trend for performance as a function of block progression is again clear (correlation coefficient of -0.57 significant at $p<0.01$). We also confirmed that the average internal noise estimate from this cohort (open circle in Figure 3) fell within the range spanned by individual estimates from the first cohort (red symbols in Figure 3).

3.7 Disengagement from decreased exploration has little impact on noise estimates

There are at least two scenarios under which test animals may display behaviour that is increasingly decoupled from the stimulus as reflected in Figure 4A-B. Under one scenario, they may swtich from stimulus-driven behaviour to free exploratory behaviour; the associated overall behavioural activity (e.g. distance travelled per unit time) may increase under these conditions, as the animals would be less and less 'locked' into maintaining their location within close range of the high-contrast stimulus. Within the context of the SDT model, this scenario would correspond to increased internal noise: behaviour becomes more 'erratic'. Under a different (in a sense opposite) scenario, test animals may reduce their overall activity altogether; this would also result in reduced behavioural drive towards the high-contrast stimulus, but it would not necessarily involve noisier behaviour.

Figure 4C plots activity (as a fraction of overall mean) across block duration (for SNR=6, the condition for which we have the largest dataset). There is a clear negative trend (correlation coefficient of -0.9, $p< 10^{-7}$) consistent with the second scenario outlined above: test animals display progressively reduced exploration of the tank (whether stimulus-driven or otherwise). This phenomenon is measurable at the level of individual experiments (distribution of correlation coefficients from separate test blocks (inset to Figure 4C) is clearly shifted towards negative values, $p< 10^{-5}$). To understand the potential impact (or lack thereof) of this nonstationary behaviour (and the associated change in binary choice exposed by Figure 4A-B) on our estimates of internal noise, we

attempted to compute separate estimates for different time epochs of each block. This is only possible to a limited extent: in order to compute percent agreement for a given trial n, we need to pool data from trial 10+n when the same stimulus was double passed. This means that the resulting estimate refers to a time window spanning half the block. Nevertheless, we can repeat this procedure for n=1, n=2, and so on. By doing this, we are effectively sliding the time window towards later sections of the block, providing at least an approximate view of how our estimates may be affected by the type of nonstationary behaviour documented in Figure 4A-C.

Figure 4D demonstrates that there was little impact of nonstationary exploration on the resulting estimates of internal noise: all except one estimate fall within the plausible range (indicated by orange horizontal dashed lines), and there was no obvious trend with time (correlation coefficient (-0.25) is not significant at p=0.5). This result indicates that the internal noise estimates generated by our protocols are to some extend decoupled from other aspects of the animal's behaviour, in the sense that they remain stable despite strong systematic changes in macroscopic features of how the animal navigates the tank. This outcome is consistent with the established finding that internal noise estimates do not correlate with sensitivity (d′), even for large datasets that support detection of small correlations (Neri (2010a, 2015)).

4 DISCUSSION

4.1 Relationship to previous studies of intra-individual variability

The measurements reported in this study represent an attempt to quantify behavioural internal noise in a non-human species within a unified theoretical framework. Internal noise is arguably the most prominent feature of animal behaviour that generalizes across sensory domains and cognitive operations (Green 1964; Dinstein *et al.* 2015). The applicability and relevance of the notion of behavioural inconsistency to animal cognition has been extensively appreciated in the literature and has been studied on multiple occasions in the form of intra-individual variability (IIV; MacDonald *et al.* 2006), however measurements of IIV have never been referred back to a normative theoretical framework that would allow quantification using the same units across different species, stimuli and task specifications. For this reason, even if IIV has been quantified for some vertebrate and invertebrate species in relation to specific tasks (Highcock & Carter 2014; Jandt *et al.* 2014), it has proven difficult to study the significance of those measurements across species.

The distinction between IIV and internal noise is made clearer by considering the fundamental methodological differences that set these two approaches apart. In typical studies of IIV, the animal

is placed within what are assumed to be identical environmental conditions, and its behavioural variability with respect to a specific trait is measured. As noted by previous authors (Highcock & Carter 2014), the assumption of an identically stable environment is in itself problematic, particularly for studies carried out in the wild: if the environment is actually changing substantially (Jandt *et al.* 2014), it may drive behavioural variability to a measurable extent. It then becomes impossible to disentangle the impact of external from internal factors onto the trait of interest. Studies of IIV not only eschew the deliberate introduction of external variability, but also do not take into account whatever variability may be intrinsic to the experimental setting (Highcock & Carter 2014; Jandt *et al.* 2014). The approach adopted in this study relies on precisely opposite premises: noise is deliberately injected into the environmental stimulus and its characteristics are finely controlled on a trial-by-trial basis to enable quantitative definition of the residual internally-driven behavioural variability. Indeed, in the absence of external modulation (the condition SNR=∞ corresponding to gray data in Figure 2) this approach is undefined and becomes inapplicable, the opposite of IIV measurements.

The approach adopted here relies on a double-pass methodology (Burgess & Colborne 1988) that is potentially applicable across a very wide range of sensory domains, task specificiations (Neri 2010a), and even species as we demonstrate here. The underlying structure and principles of the methodology remain identical, and can be referred back to the same general theoretical construct for capturing animal sensory discrimination (Diependaele *et al.* 2012): signal detection theory (Green & Swets 1966). Within this framework, internal noise is estimated in units of the external perturbation introduced by the stimulus at the level of its perceptual representation; the latter concept is applicable to zebrafish just as it is applicable to humans, or any other animal for that matter, provided it can be shown that it returns sensible and interpretable results in both cases. We have demonstrated that it is possible to identify protocols that will deliver sensible results, however our investigation has also highlighted a number of difficulties associated with this programme for future investigation (see below).

4.2 Methodological challenges

The first challenge we encountered in driving preference using our stimuli is that not all choices of stimulus specification led to useful/interpretable results. With relation to our experimental setup, preference is driven by the differences we introduce between the two stimuli presented on opposite sides of the tank. These differences are controlled by two properties: the difference between the means of the two contrast distributions associated with the two stimuli, and the common standard

deviation of those two distributions (Figure 1A-D). Stimulus discriminability or signal-to-noise ratio (SNR) is defined by the ratio of these two properties. The smallest value we tested in this study was 4 (this is not in general a small value by psychophysical standards); the two stimuli associated with this SNR level are discriminable upon cursory inspection by a human (see Figure 1A), but they often generated uninterpretable estimates of percent agreement (i.e. below chance, see black circle and lower triangle in Figure 2A). Our most reliable and useful results were delivered by a slightly higher SNR value of 6 (red data points in Figure 2). Larger values (e.g. 12) did not generate robust results (see blue data points in Figure 2); this is expected because, in the limit of SNR=∞ (gray data points in Figure 2), our methodology is not defined and internal noise estimates cannot be obtained. Intuitively, the reason for this failure is that, at very high stimulus SNR's, the externally applied noise perturbation becomes irrelevant and does not contribute to the animal's drive. Because internal noise is defined and estimated in units of external noise drive, the double-pass approach becomes inapplicable and is bound to fail. Interestingly, this is the typical regime of operation for studies relying on IIV (MacDonald *et al.* 2006; Highcock & Carter 2014).

It may seem surprising that stimulus effectiveness did not vary monotonically with SNR: for example, why should the SNR value of 6 work better than values that are both greater *and* smaller? Based on SDT considerations, we expect that large SNR values should not be viable, but we also expect that the lower the SNR value, the greater the contribution of external noise, and therefore the more effective the stimulus for internal noise estimation. Indeed, based on SDT considerations alone, a stimulus that only contains noise and no signal should be ideally suited to these experiments. The above considerations are based on the assumption that the behaviour displayed by the animal conforms to our expectations from SDT. There are many alternative scenarios, however. Consider for example the following possibility: that zebrafish may interpret excessive contrast heterogeneity (different icons taking on very different contrast values) as reflecting a non-cohesive shoal where shoal members occupy distant depth planes, and excessive contrast homogeneity (all icons taking the same contrast value) as implausible with unnatural appearance. Under this scenario, stimuli dominated by noise (low SNR) would become less attractive and would drive less shoaling; stimuli dominated by the signal (high SNR) would also drive less shoaling, but for different reasons. The end result in terms of shoaling behaviour as a function of SNR would be difficult to predict and may be non-monotonic.

We are not suggesting that zebrafish in our experiments were 'interpreting' stimuli as described above, rather we are merely offering one of many example scenarios to illustrate the notion that it would be simplistic to assume that we can predict how the animals will behave as we vary stimulus parameters, because our predictions are based on our own projected model of how the animals

'should' behave. We must first determine the way in which the animals *actually* behave; if we then wish to model specific aspects of the observed behaviour, this can only be done within the restricted range for which our model provides a reasonable approximation. In our experiments, for reasons that remain partially unclear at this stage, a stimulus SNR of $\sim$6 was able to engage the animals with sufficient efficacy to deliver reasonable estimates which cannot be attributed to stimulus-decoupled behaviour (Figure 2B) and that largely conform to SDT.

Even after suitable stimulus specifications have been identified that are effective in driving preference on the part of the test animal, there is an additional challenge associated with the deteriorating quality of such drive over time. As we have demonstrated by analyzing the progression of preference within our 14-minute blocks of 20 trials, behavioural drive steadily declines during testing (Figure 4B) and can reach chance performance within $\sim$10 minutes depending on the specific animal being tested (Figure 4A). This is not overly concerning for standard protocols where only one binary choice is measured in response to a single presentation of two competing stimuli (Engeszer *et al.* 2004; Neri 2012): 10-15 minutes are sufficient to obtain one estimate of preference using methods analogous to those used here. Application of the double-pass methodology, however, requires multiple estimates from several distinct trials during which different noise samples are delivered to the animal (Burgess & Colborne 1988; Neri 2010a). For human estimates, each block typically consists of 100 trials, each trial lasting less than 1 second. With zebrafish, estimation of preference via visually-guided spontaneous shoaling requires a longer time window, allowing us to administer only 20 trials per block. This is an important limitation of the present approach, because the SDT model underlying internal noise estimation does not incorporate the kind of non-stationary behaviour exhibited by zebrafish for spontaneous preference. It is possible that this limitation may be overcome by measuring preference under conditions of re-enforced choice behaviour, where animals would be actively rewarded for selecting the high-contrast stimulus via food delivery. Test animals may maintain more stable drive under those conditions, further enabling double-pass measurements.

Despite the drawback discussed above, the methodology proposed in this study retains a level of feasibility not afforded by other techniques. An alternative method commonly adopted in human psychophysics is the equivalent noise paradigm (Burgess *et al.* 1981; Legge *et al.* 1987). This method, however, relies on threshold measurements, each of which requires characterization of a full psychometric curve; in addition, several threshold estimates are necessary to recover the full threshold-versus-contrast function that is then used for the purpose of obtaining a single internal noise estimate. The number of trials involved is simply prohibitive for application to the zebrafish. Furthermore, the equivalent noise paradigm offers less flexibility with respect to stimulus design, which may explain why it has been used almost exclusively in visual experiments (Lu & Dosher

2008). The double-pass method is versatile, and has been succesfully applied not only to visual but also auditory phenomena (Joosten & Neri 2012). Furthermore, as explained earlier, internal noise as defined and measured in this study represents a potentially powerful tool for comparative analysis.

4.3 What is being measured by our protocol?

Behavioural measurements of intrinsic noise necessarily accumulate different sources of variability: distal transduction noise, more proximal neural noise, noise associated with the discrimination mechanism in the brain, fluctuations in motivation and attention, motor noise, and possibly others (Faisal *et al.* 2008; Dinstein *et al.* 2015). The relative weights of these different components may vary between tasks, as well as between experiments in the same task. It is also reasonable to expect that they would vary between species, not least because some components may only be present in some species and not others. Isolating the different components is a complex goal for any experimental paradigm/protocol. Because our measurements represent a first step in the direction of tackling this complex problem, it would be unrealistically ambitious to expect that the above issue would be fully resolved by this first exploratory step. Nevertheless, we believe specific features of our dataset provide at least an indication that our estimates are not confounded by certain changes in behavioural activity, and are therefore robust with respect to those changes (see below). Furthermore, although the quantitative measurements returned by our protocols may present interpretational difficulties with relation to their absolute values, they nevertheless enable conclusions based on relative changes associated with specific treatments/manipulations.

We observed clear signatures of nonstationary behaviour unfolding over the duration of each experiment ($\sim$15 minutes). With relation to binary choice, these effects are most clearly visible as a decrease in the percentage of high-contrast choices over time (Figure 4A-B). Based on more detailed analysis of the animal's exploratory behaviour (Figure 4C), we propose that this result is a byproduct of a systematic trend towards reduced exploration. More specifically, we observed a 20-30% reduction in behavioural activity over the course of the 20-trial block. It is unclear why the animal progressively reduces its engagement in this manner, but we note that it is not necessarily the case that such nonstationary behaviour should impact our estimates of internal noise as defined within the context of the SDT model outlined earlier. For example, disengagement may reflect poorer separation between the internal representations of the two stimuli, i.e. a decrease in internally represented SNR. The associated decrease in percentage of correct responses (Figure 4A-B) would be accompanied by changes in percent agreement that may allow recovery of the internal noise component in the face of the SNR changes. Our results indicate that this scenario may be applicable to our protocol and

dataset: internal noise estimates were stable across the duration of each block (Figure 4D), despite the nonstationary behaviour we documented over a similar time window (Figure 4C). Although this result does not allow us to pinpoint every component of behavioural variability that may have contributed (or not) to our estimates, it does provide evidence that those estimates did not include one clearly measurable source of behavioural nonstationarity. Future experiments will be necessary to dissect the contribution of different sources in greater detail. Our findings offer a starting point for those investigations, together with specified protocols for maximizing successful estimation.

Although as detailed above we cannot fully dissect the different sources that may have contributed to the aggregate internal noise estimates returned by our behavioural protocols, those estimates can be exploited to support conclusions about the impact of specific manipulations, such as drug delivery or targeted brain lesions. It is conceivable, for example, that specific drugs may reduce or enhance behavioural consistency (Epstein *et al.* 2011). Internal noise as assessed by our protocols may be sensitive to the effects of such agents, possibly under conditions where other behavioural metrics may not expose those effects. As we have demonstrated in Figure 4, internal noise estimates can be to a large extent decoupled from other markers of behaviour (see also Neri (2010a)), therefore potentially providing additional and complementary tools for more detailed and richer accounts of how targeted manipulations may impact behaviour. This approach would rely not on the absolute value of those estimates, but on the differential effect observed under manipulation; the latter effect would retain a significance of its own, at least as an early indicator of relevant manipulations, despite the potential difficulties associated with a full interpretation of the absolute estimated values.

When discussing intrinsic noise, the terms 'additive' and 'multiplicative' are often adopted to label different types of internal variability. This terminology can be misleading, however, because the same source of behavioural variability may be incorporated as additive or multiplicative by different models, so that model architecture becomes critical for drawing the additive/multiplicative distinction. In the standard SDT model adopted here, noise consists of a Gaussian fluctuation added to the decisional variable. In this sense, it is late additive ('late' refers to the stage at which it is added, this being the last stage before producing a behavioural response). Its unit, however, is the standard deviation of the distribution taken by the decisional variable as a result of *external* stimulus noise (this is also how d′ is defined): its intensity is therefore defined as a multiple of external noise, potentially generating confusion. For example, if external noise is varied and the estimated value of internal noise remains unchanged (as is typically found (Neri 2010b)), this means that the intensity of internal noise has actually changed and it has done so in a manner that scales proportionally with external noise by the same constant value. More importantly, because internal noise as defined and estimated here potentially encompasses multiple sources of internal variability (see above), it is

not possible to know with certainty whether its physiological origin is additive or multiplicative in nature. Our choice of model is motivated by extensive literature justifying its general applicability to human vision (Burgess & Colborne 1988; Neri 2010a, 2013).

4.4 Comparison with human estimates

Based on their bimodal distribution (histogram to the right of Figure 3), internal noise estimates from zebrafish appear to fall into one of two categories: those outside the plausible range (>5), and those within a range comparable to existing estimates from humans. This result indicates that the zebrafish may serve as a non-human model of behavioural internal noise in humans, potentially enabling a novel approach to this fundamental aspect of sensory processing. As mentioned above, internal noise may be under the control of available pharmacological agents and/or genetic factors, a possibility that could be feasibly explored in the zebrafish and subsequently transferred to human experiments (Norton & Bally-Cuif 2010). Because the trait of interest is ultimately behavioural, and because such traits may be relevant to specific pathological conditions in humans (Gilden & Hancock 2007; Perry *et al.* 2010; Kofler *et al.* 2013; Dinstein *et al.* 2015), a programme of this kind must rely on a behavioural metric supported by established interpretational frameworks and immediate generalizability across species. We propose that the class of measurements reported in this study, together with the associated experimental protocols and analytical tools, should serve as a viable candidate for future efforts in those directions. Clearly, far more data than presented here will be necessary to consolidate these tools. Our study represents only a first exploratory step in the direction of identifying whether the proposed tools may be worth pursuing in future research.

5 FUNDING

Supported by Royal Society of London (University Research Fellowship), Medical Research Council (New Investigator Research Grant) and CNRS.

References

Biro, P. A., & Adriaenssens, B. 2013. Predictability as a personality trait: consistent differences in intraindividual behavioral variation. *Am. Nat.*, **182**(5), 621–629.

Burgess, A. E., & Colborne, B. 1988. Visual signal detection. IV. Observer inconsistency. *J Opt Soc Am A*, **5**(Apr), 617–627.

Burgess, A. E., Wagner, R. F., Jennings, R. J., & Barlow, H. B. 1981. Efficiency of human visual signal discrimination. *Science*, **214**(Oct), 93–94.

Diependaele, K., Brysbaert, M., & Neri, P. 2012. How Noisy is Lexical Decision? *Front Psychol*, **3**, 348.

Dinstein, I., Heeger, D. J., Lorenzi, L., Minshew, N. J., Malach, R., & Behrmann, M. 2012. Unreliable evoked responses in autism. *Neuron*, **75**(6), 981–991.

Dinstein, I., Heeger, D. J., & Behrmann, M. 2015. Neural variability: friend or foe? *Trends Cogn. Sci.*, **19**(6), 322–328.

Engeszer, R. E., Ryan, M. J., & Parichy, D. M. 2004. Learned social preference in zebrafish. *Curr. Biol.*, **14**(May), 881–884.

Epstein, J. N., Brinkman, W. B., Froehlich, T., Langberg, J. M., Narad, M. E., Antonini, T. N., Shiels, K., Simon, J. O., & Altaye, M. 2011. Effects of stimulant medication, incentives, and event rate on reaction time variability in children with ADHD. *Neuropsychopharmacology*, **36**(5), 1060–1072.

Faisal, A. A., Selen, L. P., & Wolpert, D. M. 2008. Noise in the nervous system. *Nat. Rev. Neurosci.*, **9**(4), 292–303.

Gilden, D. L., & Hancock, H. 2007. Response variability in attention-deficit disorders. *Psychol Sci*, **18**(9), 796–802.

Green, D. M. 1964. Consistency of auditory detection judgments. *Psychol Rev*, **71**(Sep), 392–407.

Green, D. M., & Swets, J. A. 1966. *Signal Detection Theory and Psychophysics*. New York: Wiley.

Highcock, L., & Carter, A. J. 2014. Intraindividual variability of boldness is repeatable across contexts in a wild lizard. *PLoS ONE*, **9**(4), e95179.

Jandt, J. M., Bengston, S., Pinter-Wollman, N., Pruitt, J. N., Raine, N. E., Dornhaus, A., & Sih, A. 2014. Behavioural syndromes and social insects: personality at multiple levels. *Biol Rev Camb Philos Soc*, **89**(1), 48–67.

Joosten, E. R. M., & Neri, P. 2012. Human pitch detectors are tuned on a fine scale, but accessed on a coarse scale". *Biological Cybernetics*, **106**, 465–482.

Kofler, M. J., Rapport, M. D., Sarver, D. E., Raiker, J. S., Orban, S. A., Friedman, L. M., & Kolomeyer, E. G. 2013. Reaction time variability in ADHD: a meta-analytic review of 319 studies. *Clin Psychol Rev*, **33**(6), 795–811.

Legge, G. E., Kersten, D., & Burgess, A. E. 1987. Contrast discrimination in noise. *J Opt Soc Am A*, **4**(Feb), 391–404.

Lu, Z. L., & Dosher, B. A. 2008. Characterizing observers using external noise and observer models: assessing internal representations with external noise. *Psychol Rev*, **115**(1), 44–82.

MacDonald, S. W., Nyberg, L., & Backman, L. 2006. Intra-individual variability in behavior: links to brain structure, neurotransmission and neuronal activity. *Trends Neurosci.*, **29**(8), 474–480.

McIvor, A. 2000. Background subtraction techniques. *Proc. Image Video Comput.*, **6**(Apr), 147–153.

Miller, N. Y., & Gerlai, R. 2011. Shoaling in zebrafish: what we don't know. *Rev Neurosci*, **22**, 17–25.

Neri, P. 2010a. How inherently noisy is human sensory processing? *Psychon Bull Rev*, **17**(Dec), 802–808.

Neri, P. 2010b. Visual detection under uncertainty operates via an early static, not late dynamic, non-linearity. *Front Comput Neurosci*, **4**, 151.

Neri, P. 2012. Feature binding in zebrafish. *Animal Behaviour*, **84**, 485–493.

Neri, P. 2013. The statistical distribution of noisy transmission in human sensors. *J Neural Eng*, **10**(1), 016014.

Neri, P. 2015. The elementary operations of human vision are not reducible to template matching. *PLoS Comput. Biol.*, **11**(11), e1004499.

Norton, W., & Bally-Cuif, L. 2010. Adult zebrafish as a model organism for behavioural genetics. *BMC Neurosci*, **11**, 90.

716 Orger, M. B., Smear, M. C., Anstis, S. M., & Baier, H. 2000. Perception of Fourier and non-Fourier
717 motion by larval zebrafish. *Nat. Neurosci.*, **3**(Nov), 1128–1133.

718 Perry, G. M., Sagvolden, T., & Faraone, S. V. 2010. Intraindividual variability (IIV) in an animal
719 model of ADHD - the Spontaneously Hypertensive Rat. *Behav Brain Funct*, **6**, 56.

720 Saverino, C., & Gerlai, R. 2008. The social zebrafish: behavioral responses to conspecific, het-
721 erospecific, and computer animated fish. *Behav. Brain Res.*, **191**(Aug), 77–87.

722 Simmons, D. R., Robertson, A. E., McKay, L. S., Toal, E., McAleer, P., & Pollick, F. E. 2009.
723 Vision in autism spectrum disorders. *Vision Res.*, **49**(22), 2705–2739.

724 Therapontos, C., & Vargesson, N. 2010. Zebrafish notch signalling pathway mutants exhibit trunk
725 vessel patterning anomalies that are secondary to somite misregulation. *Dev. Dyn.*, **239**(10),
726 2761–2768.

727 Vargesson, N. 2007. The Zebrafish. *Manual of Animal Technology (Barnett SW ed), Oxford:*
728 *Wiley-Blackwell*, 78–84.

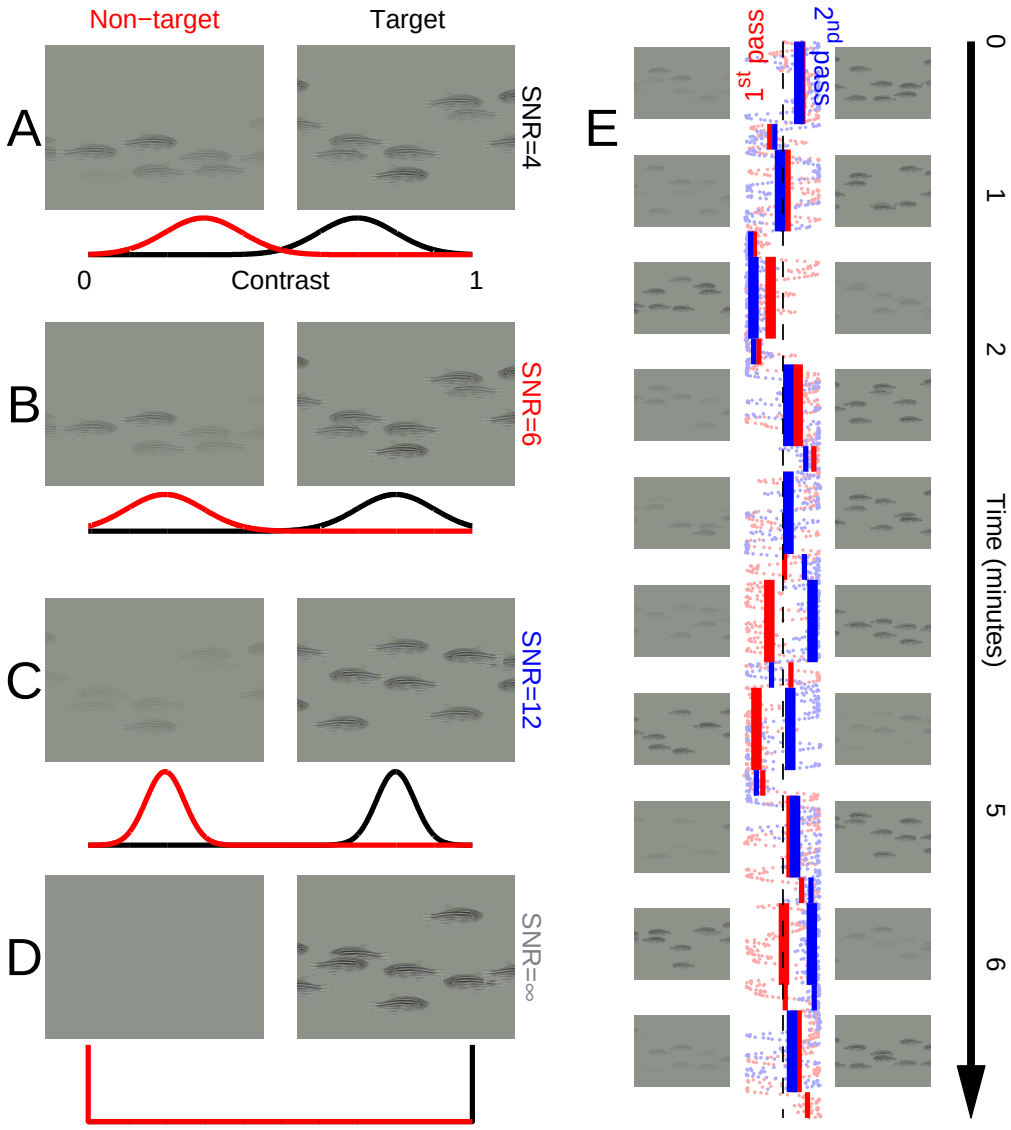

Figure 1: Double-pass procedure for measuring behavioural inconsistency. Test animals were presented with two movies of synthetic conspecifics on opposite sides of the tank, depicted by left and right images in A-D. We varied both mean and standard deviation (SD) of the contrast distributions (defined by black/red functions below images) assigning contrast to each synthetic fish. The mean contrast of the non-target stimulus (left in A-D, see red distribution) was always smaller than the target stimulus (right, see black distribution). We tested 4 stimulus configurations with increasing target/non-target discriminability (smallest in A, largest in D) controlled by larger mean difference and/or smaller standard deviation (compare distributions below stimulus images progressing from A to D). Each block consisted of 2 passes of 10 trials per pass (E), for a total of 20 trials per block. The two passes were identical: the stimulus pair presented on the first trial of the first pass (trial number 1) was identical to the stimulus pair presented on the first trial of the second pass (trial number 11), and so on. Stimulus pairs are depicted by left/right images in E. The position of the test animal along the length of the tank (horizontal axis) is indicated by small dots (one dot every 1/4 second) for first and second pass separately (red and blue respectively); the corresponding mean position is indicated by long vertical segments during stimulus presentation, and by short vertical segments during pauses between trials (blank screens). Middle of the tank is indicated by vertical dashed line: fish position to the left (right) of this line indicates preference for the stimulus indicated by the left (right) icon. Close inspection of fish position across trials demonstrates that preference was similar but not identical on the two passes, with some trials (number 1, 3-5, 8-10) presenting same preference and others (number 2, 6-7) presenting opposite preference.

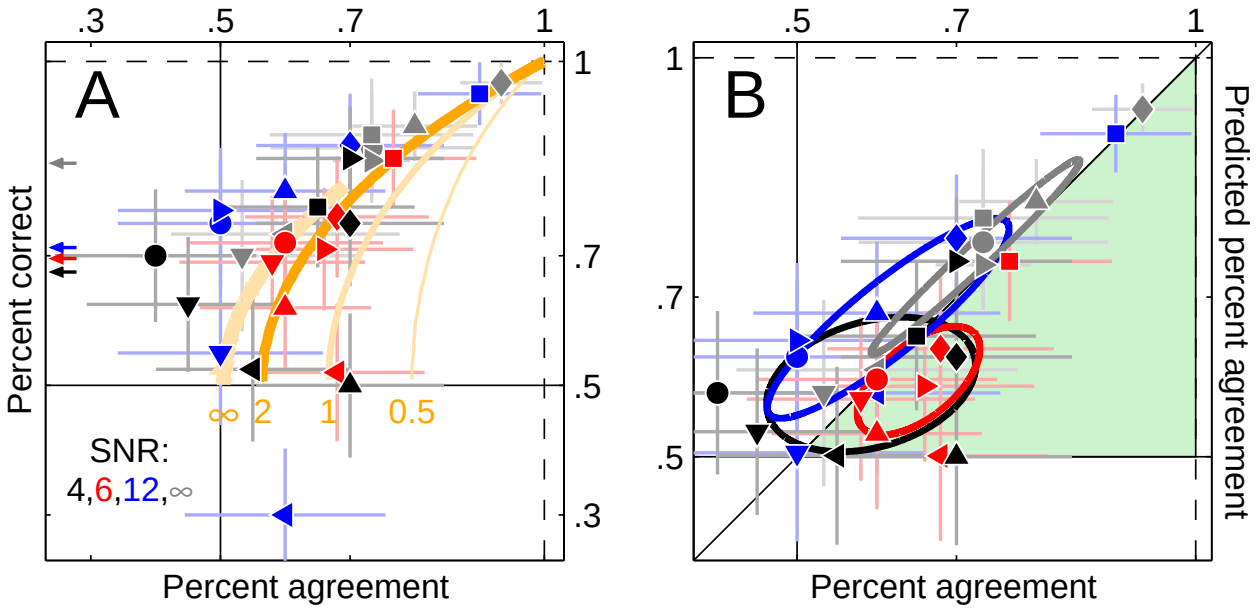

Figure 2: Zebrafish behaviour conforms to signal detection theory (SDT). The relationship between percent correct (% of trials on which the animal shows preference for the stimulus with higher mean contrast, plotted on the y axis in A) and percent agreement (% of trials on which the animal shows same preference for two identical presentations of the same stimulus pair) conforms to the predictions of SDT (indicated by orange lines, see Methods) for an internal-to-external noise ratio of ∼2 (darker orange). Percent correct demonstrates lawful dependence on stimulus discriminability or signal-to-noise ratio (SNR): the four SNR values delivered by the four stimuli in Figure 1A-D (colour-coded here by black, red, blue and gray) correspond to increasing average percent correct values (indicated by arrows pointing towards y axis in A). A certain degree of above-chance percent agreement is expected from above-chance percent correct without necessarily assuming trial-by-trial stimulus-response coupling; the y axis in B plots this expected level of percent agreement, versus the measured values (x axis, same as in A). Only the SNR=6 condition (red) is associated with empirical estimates that exceed those predicted by percent correct values alone (red data points fall below diagonal unity line within region indicated by green shading). Error bars show ±1 SEM. Different symbols refer to different (individually identified) animals.

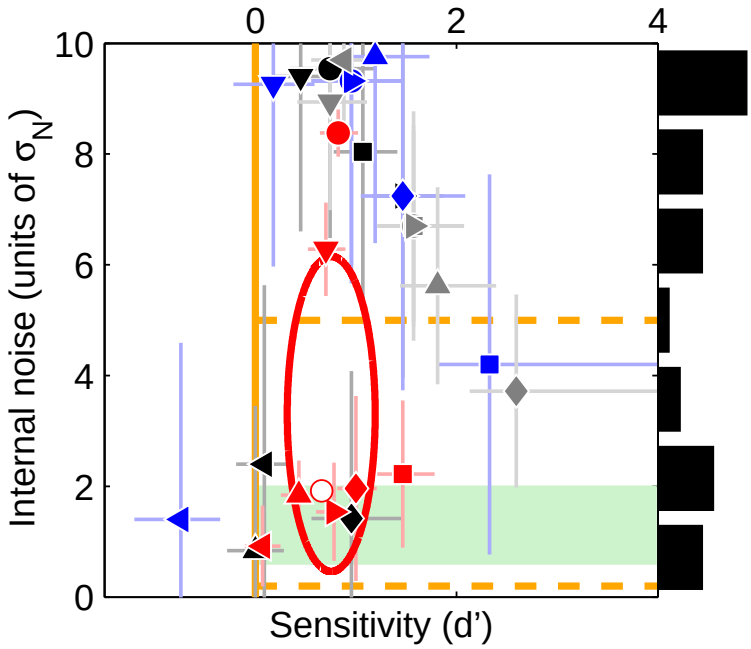

Figure 3: Direct comparison between zebrafish and human estimates of internal noise. SDT maps percent correct and percent agreement estimates (from Figure 2A) onto corresponding internal noise and sensitivity estimates (Burgess & Colborne 1988), plotted on y and x axes respectively (open circle shows average estimates across animals from the second cohort, for which individuals could not be identified separately; remaining symbols are plotted to the conventions of Figure 2). Internal noise is defined in units of external noise SD (σ_N, see Methods), sensitivity in d′ units. Internal noise estimates are bimodally distributed (histogram to the right), with roughly $1/2$ falling within the viable range (0-5, indicated by orange horizontal dashed lines) and the remaining half being implausibly large (>5). The transition point between the two groups (~ 5) is consistent with earlier work in humans (Neri 2010a), which has also identified the region defined by green shading as being representative of human internal noise. Zebrafish estimates for the SNR=6 condition (red) mostly fall within this region.

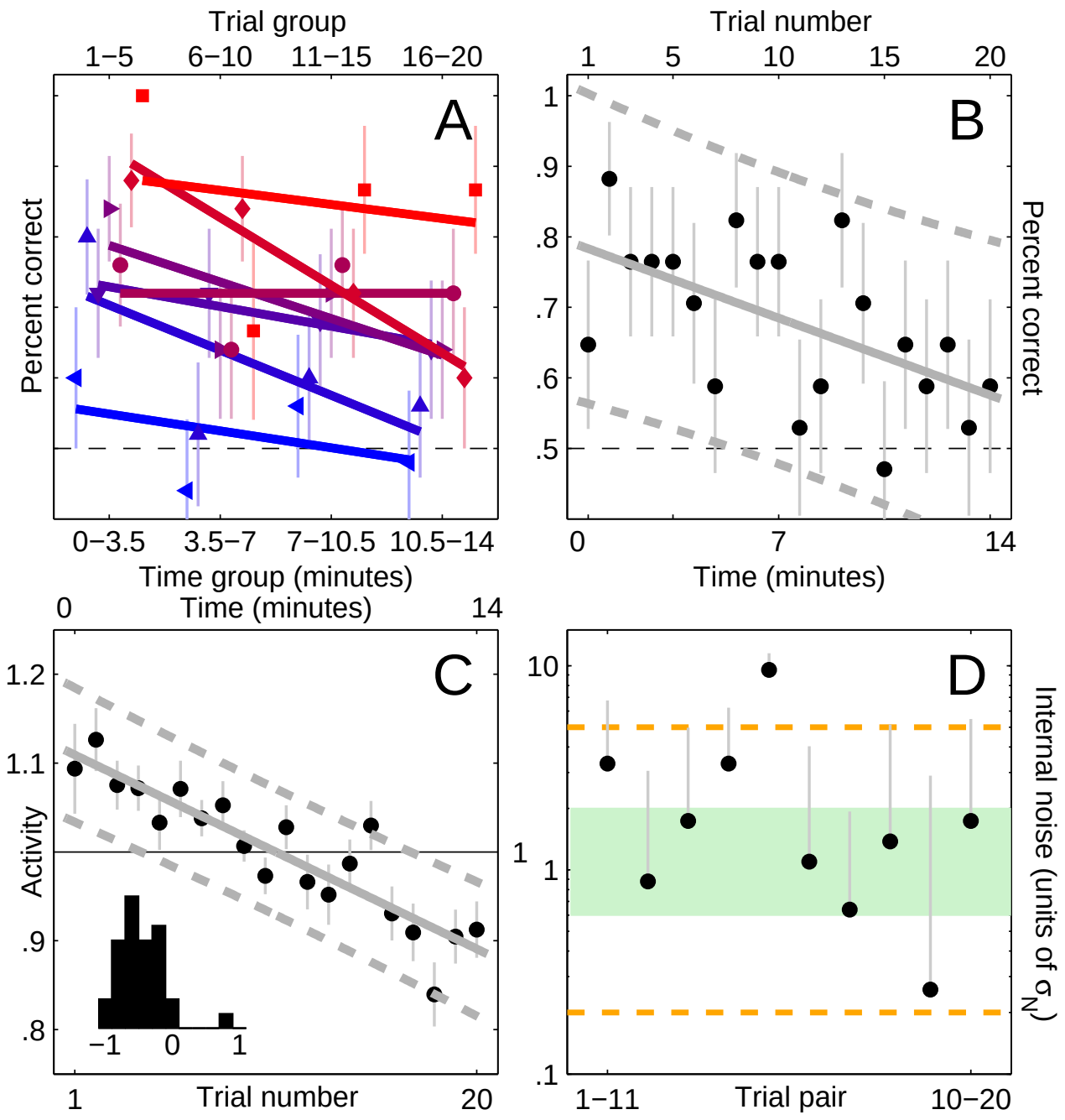

Figure 4: Shoaling preference towards high-contrast synthetic stimuli wanes during testing, but has little impact on internal noise estimates. Percent correct (y axis) is plotted in A for 4 different epochs of each test (block of 20 trials, see Figure 1E), separately for each animal (different symbol/colour). We added a small horizontal offset to data from different animals relating to the same epoch so as to avoid clutter in the plot. Lines show linear fits. B plots percent correct on each of 20 trials within a block; each value is the average across 20 animals. C plots activity (on y axis) defined as the distance travelled by the animal per unit time, as a fraction of its average value over the entire block (value of 1 means equal to average). D plots internal noise estimates from a sliding temporal window (different double-passed trial pairs, see main text) across each block; orange lines and green shaded area correspond to those in Figure 3. C-D show data from condition SNR=6 (labelled red in Figures 2-3) averaged across animals. Inset to C shows distribution of correlation coefficients for the trend shown in the main panel when computed separately for each animal/test. Solid line in B-C shows linear fit, dashed lines ±95% confidence intervals around fit.