Laterality indices consensus initiative (LICI): A Delphi expert survey report on recommendations to record, assess, and report asymmetry in human behavioural and brain research

Vingerhoets, Guy; Verhelst, Helena; Gerrits, Robin; Badcock, Nicholas A.; Bishop, Dorothy V.M.; Carey, David; Flindall, Jason; Grimshaw, Gina; Harris, Lauren J.; Hausmann, Markus; Hirnstein, Marco; Jäncke, Lutz; Joliot, Mark; Specht, Karsten ; Westerhausen, Rene

Laterality: Asymmetries of Body, Brain and Cognition

DOI:
10.1080/1357650X.2023.2199963

Published: 01/05/2023

Peer reviewed version

Laterality indices consensus initiative (LICI): A Delphi expert survey report on recommendations to record, assess, and report asymmetry in human behavioural and brain research

Guy Vingerhoets[1], Helena Verhelst[1], Robin Gerrits[1], Nicholas Badcock[2], Dorothy V. M. Bishop[3], David Carey[4], Jason Flindall[5], Gina Grimshaw[6], Lauren J. Harris[7], Markus Hausmann[8], Marco Hirnstein[9], Lutz Jäncke[10], Marc Joliot[11], Karsten Specht[9], René Westerhausen[12], LICI consortium[13]

[1] Department of Experimental Psychology, Ghent University, Ghent, Belgium
[2] School of Psychological Sciences, Macquarie University Centre for Reading, Sydney, Australia
[3] Department of Experimental Psychology, University of Oxford, Oxford, United Kingdom
[4] School of Human and Behavioural Sciences, Bangor University, Bangor, United Kingdom
[5] Department of Psychology, University of British Columbia, Vancouver, Canada
[6] School of Psychology, Victoria University of Wellington, Wellington, New Zealand
[7] Department of Psychology, Michigan State University, East Lansing, Michigan U.S.A
[8] Department of Psychology, Durham University, Durham, United Kingdom
[9] Department of Biological and Medical Psychology, University of Bergen, Norway
[10] Department of Neuropsychology, Institute of Psychology, University of Zürich, Zürich, Switzerland
[11] Groupe d'Imagerie Neurofonctionelle, CEA, University of Bordeaux, Bordeaux, France
[12] Department of Psychology, University of Oslo, Oslo, Norway
[13] List of contributing experts who consented to be named is attached below

Authors' contributions
GV initiated and supervised the LICI project, served as overviewing moderator, wrote the introduction, methods and general discussion sections, and edited the manuscript.
HV and RG programmed the survey, managed output, and wrote parts of the manuscript. Both authors contributed equally to the manuscript and qualify as second author.
NB, DB, DC, JF, GG, LJH, MHa, MHi, LJ, MJ, KS, and RW served as section moderators and/or section discussants, and reviewed the final manuscript.
LICI consortium of experts submitted and rated the statements over consecutive rounds.

Abstract

Laterality indices (LIs) are used to quantify the left-right asymmetry of a wide range of brain and behavioural variables and to provide a single measure that is statistically convenient and seemingly easy to interpret. However, there is substantial variability in how structural and functional asymmetries are recorded, calculated, and reported, suggesting little agreement on the conditions required for its valid assessment. The present study aimed for consensus on general aspects in this context of laterality research, and more specifically within a particular method or technique (i.e., dichotic listening, visual half-field technique, performance asymmetries, preference bias reports, electrophysiological recording, functional task-related MRI, structural MRI, and functional transcranial Doppler sonography). Experts in laterality research were recruited by snowball sampling and invited to participate in a three-round online Delphi survey to evaluate consensus and stimulate discussion. In Round 0, 106 experts generated 453 statements on what they considered good practice in their field of expertise. A team of moderators organized the statements into a 295-statement survey that the experts then were asked, in Round 1, to independently assess for perceived importance and their level of support, and further reduced the survey to 241 statements that were presented again to the experts in Round 2. Based on the Round 2 input, we present a set of critically reviewed key recommendations to record, assess, and report laterality research for various methods.

Introduction

Aim of the survey

One of the more important choices in laterality research concerns the *measurement of asymmetry* and the way the measured left-right difference will be treated mathematically. Debatable choices can seriously hamper the potential value of the empirical results, yet there is little consensus regarding good practices and quite different choices are used by different researchers to address similar questions. The almost complete absence of recommendations to assess and measure 'laterality' in a standard research setting is confusing and research would be much simpler if the laterality community could agree on some general recommendations. Clearly, these recommendations should not stand in the way of novel methods or approaches, nor should they be understood as prescriptive for all studies of laterality, but they would provide a good starting point against which alternative solutions should be considered. Therefore, we set out to investigate the level of expert agreement regarding common practices of asymmetry determination and formulate basic recommendations for future research that seem acceptable to a majority of laterality researchers.

Delphi approach

The Delphi technique is in essence a series of sequential questionnaires, or 'rounds', interspersed by controlled feedback, that seek to gain the most reliable consensus of a group of experts (Bishop et al., 2016; Bryden et al., 2000; Linstone & Turoff, 1975). A classic Delphi survey starts with a set of statements and goes through a series of rounds (Boulkedid et al., 2011; Hasson et al., 2000; Powell, 2003). In each round, a panel of experts rates the statements. The larger the expert sample, the greater the potential for ideas and the greater the generation of data (Hasson et al., 2000). Depending on the scientific field and the specific question, the number of experts on a panel has generally ranged from 15 to 60 (Bishop et al., 2016; Boulkedid et al., 2011; Fiander & Burns, 1998). Feedback is given that shows how everyone's ratings compare with the rest. Items can be dropped or adjusted in relation to the feedback before the next round. The entire process is anonymised in order to promote a more democratic procedure and to prevent some participants, perhaps by virtue of their prestige or seniority, from dominating the debate. Finally, the Delphi approach can be run online, which makes it inexpensive and efficient, facilitates international collaboration, and gives participants time to respond as they find convenient. Despite these advantages, the Delphi approach has also been criticized (Powell, 2003) as it may create a false sense of objectivity to the procedure (Goodman, 1987) or the resulting consensus might offer a watered-down version of the best opinion (Sackman, 1975) and result in meaningless statements that represent the lowest common denominator (Rennie, 1981). As such, it has been claimed that the results of a Delphi survey are at best an opinion (Pill, 1971) and a useful tool to generate debate, rather than a method to reach a conclusion (Mckenna, 1994). Nevertheless, it is important to emphasize that although the Delphi approach is based for the most part on quantitative ratings, it is an iterative process in which panellists, informed of the opinions of peer experts, have the opportunity to adjust their ratings or defend their position by providing comments and references.

Methods

Open data policy

In January 2020, the initiative was preregistered on OSF https://osf.io/dp3ug
All data regarding this project (correspondence, surveys, scripts, raw survey results) can be found here: https://osf.io/3kqt2/?view_only=bbc39939adc746bd89ca65bb502b261e
The files names of the corresponding document are given in the text below. See Figure 1 for a flowchart of the LICI-project. The OSF data repository is organized accordingly.

Expert recruitment

Starting from the members of the editorial board of *Laterality: Asymmetries of Brain, Behaviour, and Cognition,* snowball sampling from September 2019 to February 2020 provided the names of 220 experts who were invited to participate (see mail #1_expert panel invitation letter.pdf). Eighteen declined (retired n=5; not an expert n=4; no time n=2, no reason given n=7), 88 did not reply, and 114 experts agreed to participate.

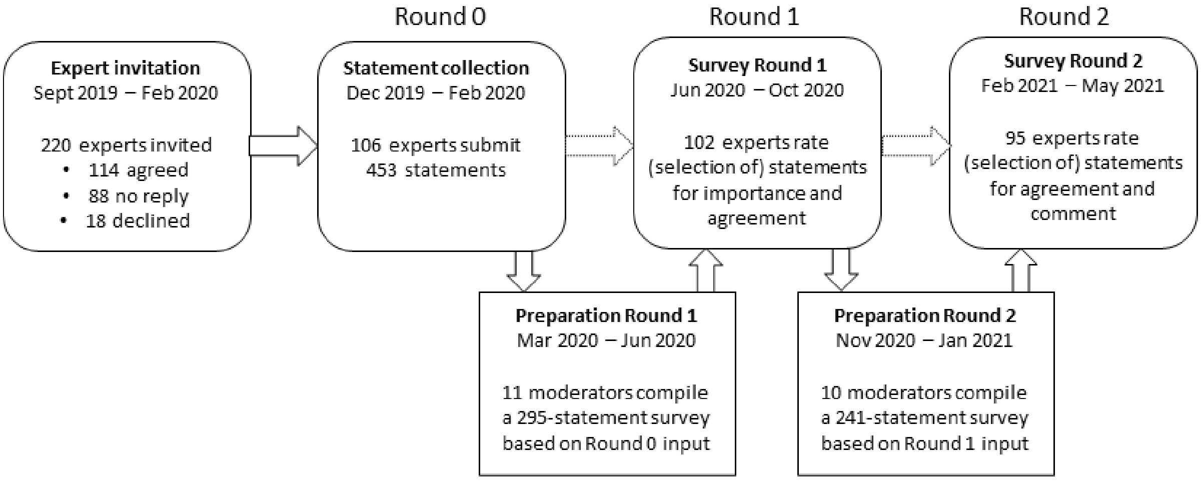
Round 0
Round 1
Round 2
Expert invitation
Sept 2019 – Feb 2020
220 experts invited
• 114 agreed
• 88 no reply
• 18 declined
Statement collection
Dec 2019 – Feb 2020
106 experts submit 453 statements
Survey Round 1
Jun 2020 – Oct 2020
102 experts rate (selection of) statements for importance and agreement
Survey Round 2
Feb 2021 – May 2021
95 experts rate (selection of) statements for agreement and comment
Preparation Round 1
Mar 2020 – Jun 2020
11 moderators compile a 295-statement survey based on Round 0 input
Preparation Round 2
Nov 2020 – Jan 2021
10 moderators compile a 241-statement survey based on Round 1 input

Figure 1. Flow chart showing participants in the Delphi survey at each round

Delphi survey

Round 0: Statement collection

As no basic material was available from which to extract statements on laterality indices, we first asked the experts to formulate recommendations they deemed important for their research. Starting from December 2019, they were invited to submit statements about their method(s) of expertise via a personal link that was included in the mail that described the procedure and provided some instructions (see mail #2_Statements invitation letter with submission link.pdf). The invitation for submitting statements was presented using the free and open source web application LimeSurvey (v3.15). A queXML schema of the Round 0 survey can be found on the OSF page (Round0_LICI.pdf). By February 29th 2020, we received a total of 453 statements contributed by 106 experts (Round 0_Statements per topic via survey.xlsx and Round 0_Statements received outside of survey.xlsx). In addition to statements on laterality indices in general, statements spanned a wide variety of methods, including dichotic listening, visual half-field technique, preference bias reports, performance asymmetries, electrophysiological recording, functional task-related MRI, structural MRI, functional connectivity approaches, and functional transcranial Doppler (Table 1). Although we inquired with some experts, no statements were received on functional near-infrared spectroscopy.

Table 1. Number of statements for each section and each round.

Section	Round 0	Round 1	Round 2*
General	86	60	44 (-20, +4, rw5)
Dichotic listening	35	18	16 (-2, +0, rw1)
Visual half-field	35	17	13 (-4, +0, rw2)
Performance asymmetries	50	57	46 (-13, +2, rw7)
Preference bias reports	75	28	27 (-6, +5, rw2)
Electrophysiological recording	25	15	18 (-0, +3, rw1)
Functional task-related MRI	62	48	37 (-11, +0, rw0)
Structural MRI	27	18	16 (-3, +1, rw2)
Functional connectivity approaches	7	-	-

Functional transcranial Doppler	51	34	24 (-10, +0, rw8)
Total	453	295	241 (-69, +15, rw28)

*Numbers between brackets indicate the number of deleted statements due to insufficient importance according to the votes of Round 1 (-), the number of added and/or split statements due to suggestions made in Round 1 and/or deemed appropriate by the moderators (+); the number of reworded Round 1 statements by the moderators to improve consensus (rw).

Moderator recruitment

To organize the incoming statements and manage their selection and eventual rewording over consecutive rounds, we set up a moderator system. Unfiltered 'Round 0' lists of statements for each section/method were reviewed by two moderators to remove duplicate or misplaced statements. One moderator served as the section expert; while the other overviewed all sections to minimize redundancy and safeguard homogeneous format and quality over the different sections. Senior experts were invited and agreed to supervise the following sections: David Carey (General section), René Westerhausen (Dichotic listening), Markus Hausmann (Visual half-field), Lauren J. Harris (Performance asymmetries), Jason Flindall (Preference bias reports), Gina Grimshaw (Electrophysiological recording), Karsten Specht (Functional task-related MRI), Lutz Jäncke (Structural MRI), Marc Joliot (Functional connectivity approaches), and Dorothy Bishop (Functional transcranial Doppler). Guy Vingerhoets served as the all-sections moderator.

By March 2020 the unorganized section lists were sent to the section moderators pre-marked for possible section misplacement or inappropriateness (too lengthy, irrelevant) by the all-sections moderator. Consensus was sought on the misplaced statements and they were assigned to another section (often the General section). Lengthy statements (sometimes spanning several paragraphs) were split into different statements where relevant and possible. Some statements having little to do with laterality indices were omitted by consensus. Remaining statements were then organized according to whether they dealt with achieving quality data, the analysis pipeline, or data reporting to help identify similar statements and to select the best-formulated. For each section, both moderators worked independently and then compared results. The final decision was again made by consensus. While most sections received at least 25 statements from multiple experts, the section on functional connectivity approaches had only 7 statements submitted by one expert. Given the insufficient number of responses to this section, both moderators agreed not to continue with this section. Although we aimed to finish this exercise on all sections by April 2020, the outbreak of the COVID-19 pandemic slowed the process, and it was only by June 2020 that we were able to send the Delphi survey out for Round 1.

Round 1: Statement selection

In this round, all experts were presented with the statements that they and their peers had submitted and that were organized by the moderators into sections and subsections. The entire survey of Round 1 consisted of 295 statements (see Table 1 for an overview of the number of statements per topic over the consecutive rounds). The survey was again presented using LimeSurvey (v3.15). A queXML schema of the Round 1 survey can be found on the OSF page (Round1_LICI.pdf). In June 2020, experts received an email with the link to the survey (email #3_Invitation Round 1 and link to survey.pdf) and were invited to rate the statements of the General section first, before proceeding to their section(s) of expertise. Ratings were based on a 5-point Likert scale. Experts were asked to rate each statement according to two questions: (1) It is important to reach consensus on the statement, and (2) I concur with the statement. Ratings for importance/agreement could be 'strongly agree / agree / neutral / disagree / strongly disagree'. A sixth option 'no opinion' was added in case participants had no expertise with this particular statement or felt that empirical evidence was lacking and preferred to abstain from voting. Round 1 was completed by October 1, 2020. See LICI_Round1_raw_cleanedup_anonymous.xlsx for the raw data results.

Round 2: Statement consensus

The results of Round 1 were screened by the all-sections moderator and distributed to the section moderators for preparation of Round 2 in November 2020. Moderators were informed that only statements deemed important by the majority of experts were to be retained for Round 2. That is, the importance rating of the combined 'strongly agree' and 'agree' ratings should be higher than 50% of the votes. Although for most statements the result of this criterion was straightforward, moderators were asked to pay attention to any statement that just failed to meet this criterion and evaluate whether its pending deletion could be due to the 'no opinion' votes and decide whether or not to retain the statement for the second round. In addition, moderators were asked to evaluate the experts' suggestions for rewording of statements or their bringing up additional statements they felt missing. We finished this process by January 2021 and prepared the Round 2 survey that was sent to the experts in February 2021 (see Table 1 for details per section).

The total number in statements of Round 2 was 241. The aim of Round 2 was to inform the experts of the results of Round 1, have them re-rate the statements based on this information, and motivate their rating (including references) or offer comment about the results in comment boxed following each statement (mail #4_Invitation Round 2 with link to the survey.pdf and Consensus Round 1.pdf). A rating procedure similar to that of Round 1 was followed. Round 2 was completed by 17th May 2021. The Round 2 survey was again presented using LimeSurvey (v3.15). See LICI_round2.pdf for a queXML schema of the Round 2 survey and LICI_Round2_raw_cleanedup_anonymous.xlsx for the raw data results. For a readable version of the Round 2 results including all the comments, consult Consensus_Round2.pdf.

A draft manuscript was prepared by Robin Gerrits, Helena Verhelst, and Guy Vingerhoets and distributed to the moderators in December 2021. Following consultation with the moderators, it was agreed to supplement each section with a critical review of the voting results and expert comments. To achieve a balanced account, we aimed for teams of two moderators per section to integrate viewpoints and reflect critically on the outcome of the survey. Because of their expertise Nicholas Badcock and Marco Hirnstein were invited to support the reviewing team.

Results

Expert panel

Snowball sampling from the initial Laterality editorial list provided the names of 220 potential experts who were invited to participate (Figure 1). One hundred and thirty-two candidate experts responded to our mail, 114 of whom agreed to participate. Compared to this initial sample, attrition over consecutive rounds was relatively low: 106 experts contributed statements (a loss of 7% compared to agreed invites), 102 completed Round 1 (a loss of 4% compared to statement contributors), and 95 completed Round 2 (a loss of 7% compared to Round 1). In Rounds 1 and 2, we collected demographic and professional information to describe the panellists. See Table 2 for the number of panellists per section on Round 1 and Round 2 and Figure 2 for an overview of the demographics per section of Round 2 panellists. A more detailed description of the expert panel for each section in Round 1 and Round 2 can be found in Supplementary Table 1. The majority of experts were located in Europe, while Asia and South-America had the least representatives, and there were no experts from Africa. Female/male ratio was remarkably balanced with a small majority for male contributors overall. While this balance is maintained over most topics, more men participated in the MRI-related methods, and more women responded to the transcranial Doppler method statements. Most experts identify with the scientific field of cognitive neuroscience and experimental psychology. The representation of clinicians was rather modest. Regarding expertise, close to 90% of the panellists had more than 10 years of experience in research, 95% had more than 10 peer-reviewed publications, and more than 75% had held their doctorate degree for at least 10 years.

Table 2. Number of participating experts for each section in Round 1 and Round 2.

Section	*N* experts Round 1	*N* experts Round 2
General	102	95
Dichotic listening	27	26
Visual half field	28	26
Performance asymmetries	46	40
Preference bias reports	42	45
Electrophysiological recording	20	17
Functional task-related MRI	28	26
Structural MRI	26	20
Functional transcranial Doppler	13	13

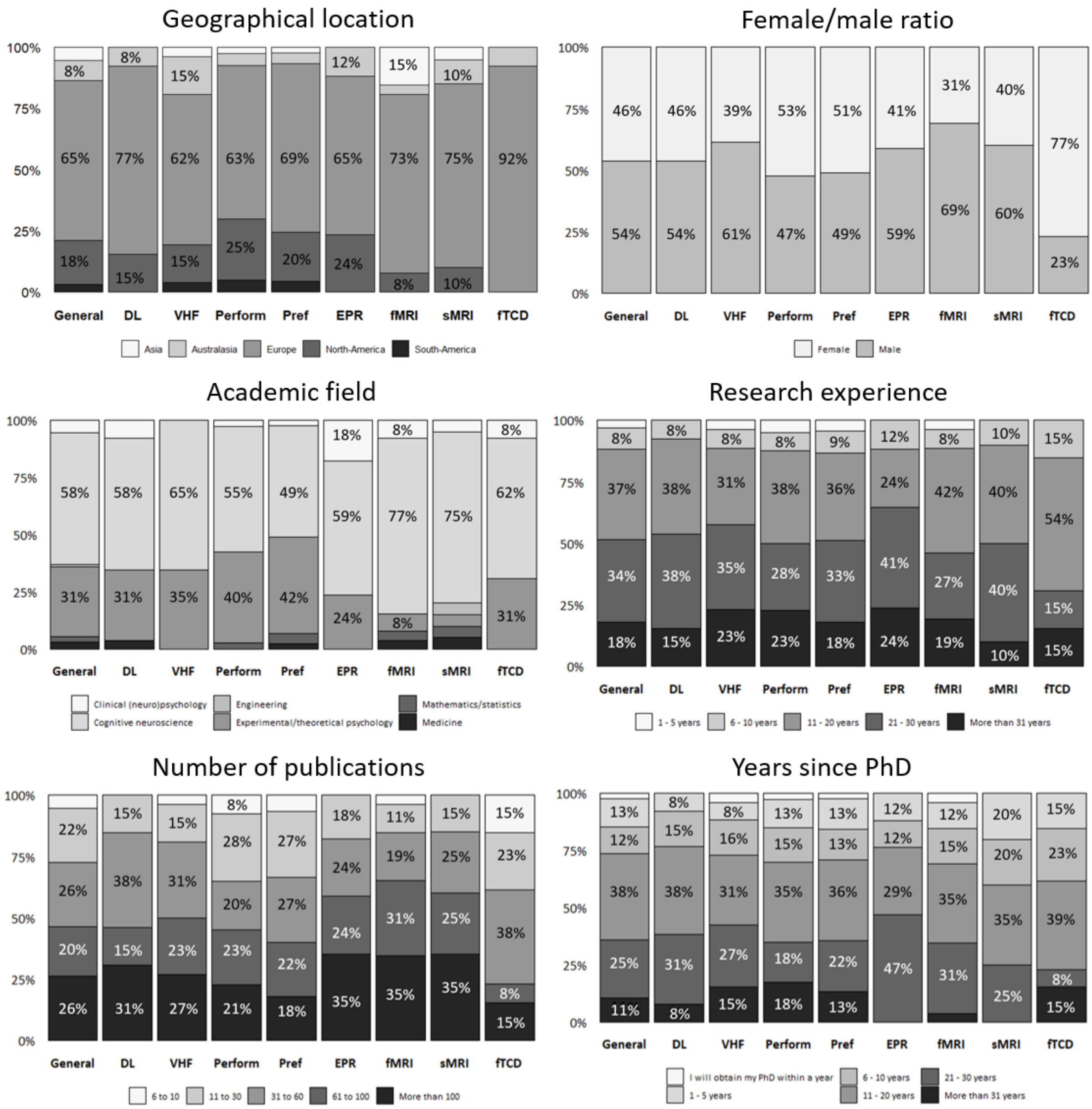

Figure 2. Overview of the panellists' characteristics per section. (DL=dichotic listening; VHF=visual half-field; Perform=performance asymmetries; Pref=preference bias reports; EPR=electrophysiological recordings; fMRI=functional MRI; sMRI=structural MRI; fTCD=functional transcranial Doppler).

Delphi-survey Round 2 results

In this part of the manuscript we present, for each section, a list of recommendations for good practice in LI-research that were able to convince a majority (>50%) of experts. Within each subsection, the recommendations are ordered according to the percentage of experts that agree (in decreasing order) by combining the 'strongly agree' and 'agree' votes into one category. The recommendations reported here often do not exactly replicate the original statement(s) but have been adapted by the moderators for clarity, style, or (most often) tone. Care has been taken to preserve the original intention of the statement. Similar statements have been combined to eliminate redundancy. The original statement(s) from which the recommendations were derived (and their percentage of agreement) are indicated between brackets. The original statements and their voting results can be found in Supplementary Tables 2-10. Each list of recommendations is followed by a critical review based on the voting results as well as on the experts' comments supplied to the statements' comment boxes, and on the general reflections and recommendations of the moderator-team that supervised the section. At the end of each section a list of outstanding issues is provided that appeared from the voting behaviour or the experts' comments.

GENERAL SECTION

Recommendations

Laterality index relevance

- Tests of motor preference should be distinguished from tests of motor performance as they represent different behavioural constructs of human motor laterality. For some research questions, there may be advantages in assessing both preference and performance, and this distinction should be acknowledged in the reporting and interpretation of scores. (GEN#8, 88.4% agreement; GEN#9, 77.9% agreement, GEN#10, 85.3% agreement).
- In general, the use of continuous measures that quantify lateralization is preferred over categorical descriptions. Where categorical classification is indicated or deemed more appropriate, categories should be clearly defined and, if derived from continuous measures, the original data should ideally be provided. (GEN#5rw, 83.1% agreement; GEN#6, 73.7% agreement).

Laterality index formulas

- Use a laterality index that reflects the left-right difference in proportion to the sum of observations rather than as a simple left minus right difference score, since a simple difference score is easily affected by the number of observations, which impedes interpretation. (GEN#12, 89.5% agreement).
- While almost 70% of experts agreed that one single preferred version of the classical laterality index would be welcome (GEN#13, 69.5%), none of the offered options (GEN#14rw) got a majority of votes. However, based on the argument that a consistent approach would be hugely beneficial to the field, the moderators suggest adopting the sixth and most popular option: The numerator of the classical laterality index should depend on the variable that is used. It is preferable to choose the numerator such that positive values (accuracy, preference) indicate a rightward bias and negative values (errors, latency) indicate a leftward bias. In this way plots are more intuitive and easier to interpret. For example the formula for accuracy measures, where higher values denote higher performance, would be (R-L)/(R+L). For reaction time measures, where higher values denote lower performance, the formula would be (L-R)/(L+R).

- A similar percentage of experts agreed that consensus regarding the scale of the laterality index would be welcomed (GEN#20, 70.9%), but again none of the offered options got a majority of the votes. To stimulate consistency in reporting, the moderators recommend the use of a proportion (range of any laterality index between -1 and +1).
- The classic laterality index has the benefit of simplicity and ease of computation, and is sufficient in many contexts. However, it is useful to be aware of alternative indices such as phi (Repp, 1977) and lambda (Bryden & Sprott, 1981), which were designed to overcome limitations of the LI. If raw data on L and R measures are provided, this would allow these alternative indices to be computed. (GEN#15, 50.51% agreement).

Reporting laterality indices

- Report background information of the sample, e.g. in humans, age, female/male ratio and handedness. (GEN#30, 97.9% agreement).
- Describe how the lateralization index is calculated (present formula and informative references) and motivate the choice of the selected laterality index. (GEN#22, 71.5% agreement; GEN#23, 91.6% agreement).
- Indicate a measure of effect size (in addition to test statistics) for all LI group comparisons. (GEN #26, 88.4% agreement).
- Where appropriate, report the 95% confidence interval (CI) of an individual's LI, rather than just the estimated value. (GEN#25, 82.1% agreement).
- If laterality is central to the research aim, also provide the raw (left and right) data in addition to the laterality index (e.g., in supplementary material or open-access repositories). (GEN#24, 79.0% agreement).
- Report individual laterality indices (whenever sample size allows). (GEN#29, 76.8% agreement).
- Present detailed descriptive statistics for data rather than just reporting statistical tests. This will vary with the context but is likely to include an estimate of central tendency (mean or median), of variation (standard deviation or interquartile range), minimum and maximum, together with an indication of whether the distribution of data is normal. For group comparisons, a measure of effect size, such as Cohen's d, should be reported (GEN #26, 74.8% agreement; also relevant to GEN#45, 74.8% agreement; GEN#28, 51.5% agreement).
- If appropriate for the method and research question, indicate whether the grand mean of a laterality index differs from zero (virtual symmetry). (GEN#31, 61.1% agreement).

Reliability and validity

- Make best efforts to provide as much evidence as possible about the reliability of laterality indices and the raw scores on which they are based. (GEN#33, 83.1% agreement; GEN#34, 58.9% agreement; GEN#35, 60.0% agreement; GEN#36, 61.0% agreement).
- Consider carefully whether the study would benefit from restricting participants to right-handers or whether both left- and right-handers should be used. (GEN#40, 50.5% agreement).

Statistical concerns

- Since the population distribution of LIs often suffers from severe non-normality, distribution-appropriate data analysis methods (e.g., for testing group differences) should be used. (GEN#45, 74.8% agreement).
- Check and report how the LI's statistical properties (e.g., reliability) differ from those of the underlying data for left and right sides. Also check and report whether task performance correlates with the polarity or magnitude of the raw data. (GEN#41, 56.9% agreement).

Laterality index calibration and decomposition

- When cut-off scores based on laterality indices are used to divide research participants into categories, or used as part of exclusion/inclusion criteria, state the rationale for using a particular score. Ideally, cut-off scores should be empirically validated and not arbitrary or only ad hoc. (GEN#58, 97.8% agreement; GEN#64, 72.6% agreement).
- Specify how handedness was defined, and use specific terms for handedness; that is, use the term 'hand preference' for bias in spontaneous choice of hand for a given task; and 'hand skill difference' for the difference in skill between the hands in a given task, with handedness as the umbrella term. (GEN#56, 89.5% agreement).
- For interpretation, use left vs. right lateralization rather than typical vs. atypical lateralization, given the unknown size of lateralization bias for a given function and a particular population. (GEN#60; 69.4% agreement).
- To improve interpretation of effects, laterality researchers are encouraged to decompose the chosen laterality index into two sub-components: direction (left/right) and absolute valence or strength (how much away from zero without +/- sign). (GEN#55rw, 65.3% agreement).

Critical review by Dorothy Bishop and David Carey

Given the large number of statements in this section we will, for the sake of brevity, limit our review to statements that were supported by the majority of experts. In addition, we will discuss statements specific to laterality indices first before turning to more general recommendations concerning reliability, reporting, and open access.

Laterality index relevance. Although LIs are considered relevant by virtually all experts, commenters warned that a single measure implies a reduction of information and that values for each side may also be crucial. The relative difference between sides is important, but the source of the relative difference may be informative to understand the mechanism at play. Comments underlined that in specific contexts LIs may be unsuitable or deceptive, and that more sophisticated models (e.g., linear mixed models) may diminish the need for a single LI for statistical purposes. Remarks also referred to the differences between registration techniques (Seghier, 2008), and the conceptualization of laterality as a continuous or categorical variable. Although LIs can be measured on a continuum, this may not be appropriate for all asymmetries (e.g., handedness; (Buenaventura Castillo et al., 2020; Busch et al., 2010; McManus et al., 2016; Tran et al., 2014).

Need for operational definitions. Clear operational definitions and guidelines on the use of laterality measures to achieve reproducibility and transparency were welcomed. Experts supported their case by references from very different topics (Adcock et al., 2003; Benjamin et al., 2017; Bradshaw et al., 2017; Carson et al., 1993; Edlin et al., 2015; Hardie & Wright, 2014; Jansen et al., 2006; Mathew et al., 2019; Ramsey et al., 2001; Sainburg & Schaefer, 2004; Seghier, 2008; Wegrzyn et al., 2019). It was, however, noted that standard guidelines might not be appropriate for all contexts and should not be required for publication. Well-justified and well-documented deviations from standard methods should continue to have their place in science.

Cross-species and cross-age comparisons. It was generally agreed that we need tools that allow testing of brain and behavioural biases across ages and species, but commenters noted this might be difficult to achieve and interpret, especially across species. There appears to be more enthusiasm for developing methods that allow for across-age testing, for example by dynamically varying stimulus exposure durations according to performance (Cherry et al., 1995).

Comments also reflected the debate over whether or not laterality should be conceived as a continuous or categorical variable. There seems little compelling evidence for either position at present. Binary measures may be more intuitive, though continuous variables have statistical advantages and allow for finer distinctions. They can also easily be converted to binary measures (Busch et al., 2010; Labache et al., 2020; Tran et al., 2014). In general, the votes favoured quantitative (continuous) measures of laterality for many research purposes, but whether continuous or binary measures are more useful will depend on the specific question(s) of interest and how the underlying phenotype is conceptualised. In clinical settings, where the surgeon may have to make decisions about unilateral brain surgery, a quantitative measure of a patient's language dominance is likely to be converted to a binary left vs right score. Comments from clinical experts described the methods (and the associated problems) used to translate neuroimaging data into surgical decisions (Baciu et al., 2005; Bradshaw et al., 2017; Branco et al., 2006; Fesl et al., 2010; Jansen et al., 2006; Nagata et al., 2001; Seghier, 2008; Suarez et al., 2010; Wegrzyn et al., 2019; Wilke & Lidzba, 2007; Wilke & Schmithorst, 2006). Outside of medical conditions, commenters expressed little enthusiasm for the use of binary or tertiary classification, unless the use of categories is clearly explained in relation to specific objectives. A final point to note is that if researchers use cut-offs to divide a laterality continuum into categories, they should be explicit as to whether this was done on a priori grounds, or after inspection of the data. The latter approach generates a high rate of false positive findings (Bishop, 1990b), and results obtained this way cannot be interpreted with confidence unless replicated in a new sample.

Dynamic aspects of laterality. Comments on a statement referring to the stability/changeability of laterality revealed that the idea is embraced by some and distrusted by others. Some experts argued that laterality could be dynamic but did not consider this relevant when studying homogeneous tasks. Others found it an interesting and understudied question and wanted to broaden it beyond temporal and spatial scales, for example by considering a hierarchy of cognitive functions in which some (e.g., language) are more dynamic than others. Importantly, the statement was backed-up with references showing empirical evidence for (very) short time variation in asymmetric brain activity (Jayasinghe et al., 2021; Mohr et al., 2005; Schaefer et al., 2009; Serrien et al., 2006). The key question is how laterality researchers should respond to this evidence. If laterality is dynamic, how do we set a standard to measure a laterality index and how do we address its temporal and spatial changes? It is suggested that while dynamics in asymmetry may be relevant for some types of measure, others may show more stable patterns of laterality. For instance, hand preference beyond early childhood is a stable trait, and test-retest reliability of language laterality on fMRI and functional transcranial Doppler ultrasound is high (Johnstone et al., 2020; Stroobant et al., 2011; Stroobant & Vingerhoets, 2001; Woodhead et al., 2018; Woodhead et al., 2021).

Preference vs. performance. The importance of distinguishing between tests of preference and tests of performance was widely acknowledged and supported by several references on motor laterality (Musalek, 2014, 2015) or studies investigating relations between preference and performance (Brown et al., 2004; Musalek et al., 2015; Rigal, 1992). At the same time, it was remarked that assessing both may not be needed for every research question. References exploring the use of composite laterality indices that merge performance and preference scores were provided (Brown et al., 2004; Bryden et al., 2000), but several experts were unfamiliar with this approach or questioned its validity.

Context-dependency. Commenters who agreed that lateral biases were context dependent, provided references that mostly came from the motor field (Bishop, 1989, 1990a; Flowers, 1975; Goble & Brown, 2007; Guiard, 1987; Hausmann et al., 2004; Perrier, 2015; Porac, 1981; Provins, 1956; Sherwood, 2014; Todor & Doane, 1977).

Laterality index formulas. Advocates for a consensus laterality index underlined that it would improve the comparison between studies, benefit meta-analytic efforts, and aid replication and interpretation. On the other hand, some comments expressed concern that requiring use of agreed measures may become dogmatic and limit scientific freedom. This illustrates the inevitable tension between achieving consensus and novel discovery. It is not just the formula, but many other methodological choices like threshold and region-of-interest selection that hamper comparability and interpretation. To summarize the comments, an agreed LI version would have advantages, but should not preclude alternative solutions if appropriate. Selection of the LI (classic or alternative) should be motivated. As expected from the comments, voting on a favourite general LI did not produce a clear winner. The range of concerns was broad. It was clear that what is intuitive to some may not be to others. Where laterality is categorical, there are questions as to how to handle cases of 'Both' and 'Either' in a formula. Some formulas may not suit particular methods and it may be difficult to find a formula that suits behavioural *and* neurological data. LIs also face statistical constraints (Verhelst et al., 2021). Some commenters reported that they found it difficult to select one of the options because they lack the comparative methodological and statistical evidence upon which a consensus could be based.

Although division between experts might lead to the conclusion that we should let researchers continue with their favourite approach, so long as it is clearly described, we (Carey and Bishop) see the current Delphi exercise as an opportunity to make a definite recommendation for researchers to adopt a consistent approach, as this would be hugely beneficial to the field. We propose the LI as specified in the sixth option of GEN #14rw, i.e., (R-L)/(R+L), which also received most votes. Although the runner up option (with the denominator described as (R+L)/2) also attracted about a quarter of the votes, its resulting LI would take a value between -2 and +2, which we feel might be counterintuitive. A measure derived this way can be expressed as a proportion or a percentage. And while only 71% agreed for the need to achieve consensus on this point, we recommend use of a proportion, to meet our goal of having consistency in reporting.

Reporting laterality. A fair number of statements were intended to serve as guidelines for adequate reporting of LIs and related information. Inclusion of background information of the sample (age, sex, and handedness) for example, is fairly standard advice for good scientific reporting, as is the recommendation to describe and motivate the selected laterality index. Researchers who use the classic LI could cite this Delphi article, justifying the adoption of that form of LI to improve consistency in the field.

While most statements on 'Reporting LIs' were acceptable to more than 70% of experts, commenters considered some as being too prescriptive, not always feasible or appropriate (e.g., measures of effect size (Grissom & Kim, 2012), or dependent on the research questions. The original version of the recommendation to provide detailed descriptive statistics perhaps sounded too prescriptive, which may explain why support was not unanimous. The motivation behind this item was to ensure that assumptions of statistical tests are met, and to make it possible to consider the possible impact of outliers and non-normality on results. In our recommendation, we have rephrased this item to enhance consensus, as it is of particular importance in laterality research, where non-normality is common. The inclusion of effect size estimates in reporting has been recommended by publication guidelines for many years and can be useful when the results are incorporated in subsequent meta-analyses.

Open data. Many of these disagreements can be resolved if raw data are provided alongside the paper as supplementary material and/or in an open access repository; the reader can then calculate whatever is deemed interesting without having to overload the manuscript with descriptive statistics or measures that are of secondary relevance to the research question. In addition, if researchers make their raw data available, then this allows other researchers to compare the impact of adopting different measurement methods, which ultimately may lead to adoption of an approach that is best-supported by evidence.

Reliability and validity. Similar remarks are made with regard to recommendations on reliability measures. While (documented) LI-reliability is deemed important by most (De Schryver et al., 2016; Fernández et al., 2003; Friedrich et al., 2017; Polit, 2014), comments also criticized the impracticality of this guideline as reliability data may not be available (especially for tasks not widely used) and grants generally do not sponsor test-retest reliability research. In addition, such measures might not be necessary in every context and there seems to be some confusion and uncertainty of what the best approach/measure would be. When it comes to setting a minimum number of items opinions are mixed. While many experts endorse the idea of a minimum requirement, it is commented that this might not always be feasible in some groups (infants), wonder how one should determine a minimal number, or outright reject the idea as being overly coercive. Nevertheless, trial numbers impact handedness assessment (Campbell et al., 2015) and the recommendation to select a standard handedness inventory, with an agreed minimal number of items to calculate handedness indices is also clearly advanced. In addition, there is growing evidence that reliability of measures is crucial for any studies focused on individual differences (Parsons et al., 2019), and the need to report reliability of measures is part of the APA Journal Article Reporting Standards (JARS). In particular, if we do not know how reliable a measure is, then we cannot tell whether any failure to distinguish two groups, or to correlate with another measure, reflects true dissociation or just poor measurement. For tests that are based on a series of items, simple split-half reliability from a correlation of LI from odd versus even items provides an indication of how reliable the measure is (using a nonparametric correlation coefficient if the data are non-normal). It is more demanding, and may not always be feasible to assess test-retest reliability, but if data exists, it should be reported, and it is advisable to conduct a test-retest study before committing resources to a full-scale study using a specific measure of unknown reliability.

Handedness of samples. A statement on handedness-based participant selection elicited many reflections and strong opinions in favour or against the use of homogeneous groups. The gist of the different viewpoints appears to be that this is strongly dependent on the goal and context of the study. If the aim is to represent the general population, a mixed sample might be preferable. But if hand preference is not central to the research question it might be convenient to select only right handers to minimize confounding variance even at the cost of limiting generalizability (Bailey et al., 2020; Willems et al., 2014). In sum, most commenters advised not to specify this as a general rule. So, while the original statement (GEN#40) specifically recommended against including only strong right-handers, it did not get strong endorsement. We propose rephrasing the statement such that participant selection criteria should be set to be optimal for the study question.

Statistical concerns. Several statements on statistical concerns of LIs were criticized for being not specific enough or for being too specific to a research context to be endorsed as general guidelines. References were provided that illustrate issues of complex interrelations (Seghier et al., 2011), non-normality (Seghier, 2019), or alternative approaches to LIs (Mathew et al., 2019; Sainburg & Schaefer, 2004). In evaluating the comments made in the ‘Reporting’ and ‘Statistics’ sections, there was a sense of tension between researchers who saw the Delphi exercise as an opportunity to increase the rigour of reporting in this field, and those who were concerned at new demands for additional methods or analyses that might act as a straitjacket for researchers, stifling creativity and making all studies uniform. In practice, many of the points made are not specific to the area of laterality, and could be subsumed under two general recommendations: (1) follow the APA JARS for quantitative research in psychology (https://www.equator-network.org/reporting-guidelines/journal-article-reporting-standards-for-quantitative-research-in-psychology-the-apa-publications-and-communications-board-task-force-report/), and (2) report results as if you are anticipating having your study incorporated in a meta-analysis in the future. Several of the comments mentioned the importance of providing data in a format that is suitable for meta-analysis, and it seems that meta-analytic studies of laterality research are becoming increasingly common–- yet have been frustrated by inconsistent and/or incomplete reporting.

Laterality index calibration and decomposition. Guidelines for LI calibration and decomposition elicit mixed feelings. Commenters agreed that while standardized boundaries may not be appropriate, cut-offs should be well justified (Seghier, 2019) and should be determined beforehand to avoid false positive results (Bishop, 1990b). Decomposition of the LI in direction (left/right) and strength is considered valid only if there is an a priori hypothesis regarding both components, otherwise a signed LI would suffice.

Outstanding issues

Is it feasible to make tests of behavioural and brain biases effective for testing across ages and perhaps even across species? (General #3).

Under what circumstances should laterality be conceived as a categorical or continuous variable? (General #5rw, General #62).

If laterality is a dynamic process, how do we address the temporal and spatial changes when measuring a laterality index? (General #6).

Despite the observation that close to 70 percent of experts would favour a single preferred version of a broadly deployable lateralization index (General #13), there is little agreement on its actual formula (General #14rw). Are there methodological or mathematical arguments in favour of or against the different options that we could articulate to come to a more informed decision? If no consensus on a single preferred LI formula can be reached, it might be an option to recommend which types of LI to use under which circumstances (decision tree).

A careful selection of the items is considered by many a crucial step of any approach for a valid quantification of laterality (General #37). Many experts also call for an agreement on the minimal number of items for handedness indices (General #38). But what is 'careful selection' and is the minimal-item requirement not too context-dependent? It may be relevant to articulate a list of recommendations about item selection and item quantity. This might be particularly useful for the determination of handedness.

For each measure, unanimous agreement on categorization would be useful. For example, experts agreed on the need for a clear definition of right-handed, left-handed, and ambidextrous (General #57). Since no formal proposal was made, this may be something for the experts on handedness to debate. Some may not believe in true ambidexterity, but if we can agree on what an ambidextrous person should be, we can determine how many individuals meet these criteria (just as we can define a unicorn without it existing), or we can propose another, more neutral, term for individuals with no consistent directional hand preference.

DICHOTIC LISTENING

Recommendations

Hearing deficits

- For DL tests, include a check to ensure headphones are the right way around. (DL#4, 100%).
- Account for occasional hearing deficits in participants of DL paradigms, and explain how this was achieved (e.g., self-report, basic non-dichotic screen, probabilistic control by large samples, audiometry, …) as well as the exclusion criteria used. (DL#1, 57.7%; DL#2, 73.1%; DL#3, 84.6%).

Reliability

- Use a sufficiently high number of trials to ensure an acceptable degree of retest reliability. (DL#5, 92.3%).

Task construction

- Dichotic stimuli might at first appear confusing to the participant. To overcome this, use practise trials before starting the test to familiarize participants with the testing situation. (DL#12, 100%).
- Maximize the spectral and temporal overlap of dichotic stimuli across channels to promote cross-channel competition and stimulus fusion. (DL#7, 75.4%).
- Include trials in which the same stimulus is presented to both ears. (DL#9rw, 73.1%).
- Unless the research question involves workload, use single dichotic stimulus presentation per trial to assess hemispheric lateralization as it reduces the working memory load compared with paradigms that use multiple dichotic stimuli per trial. (DL#8, 53.8%).

Data reporting

- Report left- and right-ear correct recall in addition to the laterality index to identify which side contributes to changes/differences in laterality. (DL#14, 92.3%).
- Determine the laterality index so that left-ear preference results in a negative value, and a right-ear preference in a positive value. (DL#15, 84.6%).
- Various measures can be collected in DL to assess perceptual laterality (e.g., the score for each ear might be number of correctly identified stimuli, reaction times, signal-detection sensitivity, disruption from noise or interference, delayed recall, etc.). Laterality indices need to consider the difference in the characteristics of these measures. (DL#17, 84.4%).
- Calculate and report multiple dependent variables, e.g., magnitude of the right-ear advantage and proportion of subjects who show a right-ear advantage. (DL#18, 76.9%).
- Use a laterality index that standardizes the interaural difference to the overall level of performance (sum of left- and right-ear correct recall). (DL#13, 69.2%).

Critical review by Markus Hausmann and René Westerhausen

Overall, there was a good consensus across laterality experts on the DL technique statements. At least two-thirds of all experts (65.4 - 100%) agreed on 12 out of 16 statements.

Hearing deficits. Experts generally agreed with the hearing-deficit statements. There was only one statement (DL#1) where the agreement was slightly less pronounced (57.7%). This statement proposing exclusion thresholds for hearing deficits was criticized because there is concern about the proposed threshold (too restrictive, unclear formulation) (Iliadou et al., 2009; Iliadou et al., 2017) and that not all researchers have access to audiometry. Commenters advanced alternative approaches (e.g., self-report, basic non-dichotic screen, probabilistic control in large samples) they believed should not be excluded. In addition to a check for proper headphone placement it is further suggested to counterbalance headphone placement to control for equipment problems. It is important to note, however, that assessment of auditory laterality is affected by hearing acuity differences between the ears, as asymmetric hearing deficits shift the laterality towards the better ear (Speaks et al., 1983; Speaks et al., 1980). Studies examining the effect of interaural-intensity difference in stimulus presentation (Berlin et al., 1972; Hugdahl et al., 2008), might additionally be taken to predict that acuity differences of about 6 to 12 dB in disfavour of the right ear, eliminate the right-ear advantage in verbal dichotic paradigms. Thus, a hearing test to assess acuity differences between ears can provide valuable additional information both to exclude participants or potentially correct auditory laterality measures, especially when the aim is to classify individuals according to their hemispheric dominance. Considering group comparisons with larger samples, one might assume that individual differences in hearing asymmetry will stochastically even out. Although also this assumption might be violated, when systematic difference between groups in hearing acuity exist (e.g., in aging research, (Passow et al., 2012); or in schizophrenia, (Kompus et al., 2013). In summary, it appears reasonable to recommend that, if hearing acuity was not tested, the reasoning behind this choice is reported.

Reliability. Almost all experts (92.3%) agreed that the number of trials should ensure an acceptable test-retest reliability which raises questions on how many trials would be enough and a reference was offered suggesting reliability starts reaching ceiling at 75 trials (Parker et al., 2021). Possible contamination of data due to repetition of stimuli which then lose novelty was also remarked. Overall, only 50% agreed that the test-retest reliability of the index score should be stated. Reporting the reliability of the index scores is questioned on the difference between experimental designs and individual differences designs and the fact that it would require many trials for each experiment inducing unnecessary fatigue that might be avoided when using established paradigms whose retest reliabilities are known. It is important to note, however, that the reliability of the used paradigm is important as it, following the classical test theory, also determines the upper limit of its validity (Pedhazur & Schmelkin, 2013). While the number of trials is an important determinant of reliability (Speaks et al., 1982; Westerhausen, 2019), it also will be affected by other design choices; for example, the stimulus material or the response format. Thus, reliability estimates of dichotic-listening paradigms are very heterogeneous (Voyer, 1998), and estimates between r = .63 (Wexler & King, 1990) and r = .93 (Westerhausen & Samuelsen, 2020) can be found even in paradigms with 120 trials. Thus, it appears difficult to make general suggestions about the minimal or maximal number of trials, without assessing reliability empirically. The question of the optimal number of trials always represents a compromise between reliability and feasibility considerations (such as fatigue, reuse of stimuli, experimental time constraints). Nevertheless, researchers should be encouraged to provide reliability estimates with their findings.

Task construction. Two out of five statements concerning task construction received less support. This referred to the statement DL#11 (50%) which proposed to consider intertrial effects such as negative priming (50%), and statement DL#8 which proposed single dichotic listening stimulus presentation as it reduces working memory load compared to multi-stimulus paradigms (53.8%). However, one commenter interpreted this finding as confusion on what is meant by single dichotic stimulus presentation: a single trial, a monaural presentation. Other commenters interpret it as the presentation of one single unit at each trial and remark that sentence-length stimuli, which can be regarded as single unit stimuli, also achieve higher memory loads. In addition, some remarked that the investigation of memory load on laterality can be a key part of the research question. One commenter stated that the statement on the selection of sample appropriate stimuli (DL#9) is another potential source of confusion, as some commenters might have interpreted it to suggest the selection of age and culture appropriate stimuli, while others see it as a manipulation geared to check the validity of the performance and identify participants who are either not motivated to respond carefully or cannot hear the differences between stimuli. The idea of maximizing stimulus fusion (DL#7) seemed generally accepted, but commenters wondered whether there is empirical evidence to back this up, and whether this recommendation is sustainable for research on the effects of temporal and spectral manipulation of dichotic stimuli.

It is long tradition is dichotic-listening literature to pair stimuli, which differ only in one phoneme, such as rhyming words (Hiscock et al., 2000; Wexler & Halwes, 1983) or CV/CVC syllables with the same vowel (Hugdahl et al., 2009; Shankweiler & Studdert-Kennedy, 1975). The aim behind this approach is to achieve spectral/temporal overlap of the two sound stimuli so that the stimuli are likely to perceptually fuse and will be subjectively perceived as one stimulus (Cutting, 1976; Repp, 1976). This has been considered beneficial as it reduces the information to be processed to a single item, which minimizes the cognitive demands during stimulus selection (Westerhausen et al., 2013; Wexler, 1988) and reduces the tasks susceptibility to voluntary attentional shifts (Asbjørnsen & Bryden, 1996). The number of stimulus-pair presentations per trial also vary between dichotic-listening paradigms, as participants have been confronted with a single pair of stimuli (Hugdahl et al., 2009; Wexler & Halwes, 1983) or with multiple pairs (Kimura, 1961; Musiek, 1983) before a response is to be given. Presenting multiple pairs per trial, increases the cognitive demands for the participant: a large number of items has to be kept in working memory (increasing memory load and resulting in "forgetting", see (Aghamollaei et al., 2013; Penner et al., 2009)) and, consequently, a subjective report strategy is developed (e.g., reporting one ear before the other; see (Bryden, 1962; Freides, 1977). Both effects have been shown to systematically modulate the magnitude of the right-ear advantage in verbal dichotic paradigms (Bryden, 1962; Penner et al., 2009). Thus, it has been argued that whenever it is the research question to assess auditory laterality, the aim should be to minimize the effect of these (and other) cognitive factors on the obtained measures (Westerhausen, 2019; Wexler, 1988). However, if the aim is to specifically assess cognitive functions, like working-memory capacity, the argumentation can be easily reversed. Thus, despite of the general agreement on the above statements in the community, it appears difficult to provide general recommendations for the design of dichotic-listening paradigms that apply for all research questions. Rather, it is recommended that authors provide reasoning behind their design choices for the dichotic paradigm.

Data reporting. Statements on data reporting, which received high agreement (>76.9%), were often accompanied by comments stating that the publication of raw data is the best strategy to accommodate these recommendations. In addition, it was remarked that dichotic listening allows for the collection of various measures to assess perceptual laterality (e.g., the score for each ear might be number of correctly identified stimuli, reaction times, signal-detection sensitivity, disruption from noise or interference, delayed recall, etc.), but that testing/reporting all these measures might not be needed if there were more empirical data on the relation between these measures. Also, it is good scientific practice to make all data publicly available (e.g., in the appendix or open science platform) and share them with the laterality community.

Outstanding issues

What evidence is there that maximizing the spectral and temporal overlap of dichotic stimuli across channels, thereby promoting cross-channel competition and stimulus fusion, gives rise to more reliable and valid laterality information? Does this evidence result in practical recommendations that can be applied by researchers when constructing their paradigms?

Various measures can be collected in dichotic listening to assess perceptual laterality (e.g., the score for each ear might be the number of correctly identified stimuli, reaction times, signal-detection sensitivity, disruption from noise or interference, delayed recall). For a better interpretation of the effects of measure selection on laterality indices in dichotic listening, it seems relevant to obtain empirical data on the relation between these measures.

VISUAL HALF FIELD

Recommendations

Eye movements

- Eye tracking is good but not necessary if other measures are taken to make sure participants fixate well (e.g., if participants are required to process information on some trials at the fixation location). (VHF#1, 92.4%).
- To avoid short saccades (<100 ms), have the fixation stimulus remain visible while the parafoveal stimuli are presented. (VHF#2, 77%).

Task construction

- If manual responses are measured, stimulus-response compatibility effects should be avoided (e.g., by switching hand during the experiment or requiring bimanual responses). (VHF#3rw, 89%).
- Keep VHF presentations of stimuli in the range of 100-180 ms (depending on task difficulty). (VHF#6, 76.9%).
- Use stimulus eccentricities of 1 degree visual angle off fixation and not exceed 6 degrees visual angle. (VHF#8, 65.4%).

Reproducibility

- For VHF studies, use a chin rest to ensure that head movements are minimised and that a constant distance from the monitor is maintained. (VHF#10, 69.2%).

Analysis pipeline

- In VHF studies, report laterality indices for both response time and accuracy measures, if possible. (VHF#14, 92.3%).
- For calculating response time-based laterality indices, include only correct responses. (VHF#15, 92.3%).
- Include a VHF laterality index to take into account the overall level of performance and not only the left-right difference. (VHF#13, 84.6%).

Data reporting

- In VHF studies, report LVF and RVF performances (means and standard deviations/errors) in addition to laterality indices because the calculation of laterality indices results in a loss of information. (VHF#17, 88.5%).

Critical review by Markus Hausmann and Marco Hirnstein

Overall, there was a good consensus across laterality experts on the VHF technique statements. At least two-thirds of all experts (65.4 - 92.3%) agreed on ten out of 14 statements. It is important to note that laterality experts see the majority of individual control measures mentioned in the present paper to be less critical. This probably makes sense as the reliability of the VHF technique depends largely on the combination of several control measures. For example, eye tracking (VHF#1) might not be necessary if experimenters present stimuli bilaterally (VHF#5), tachistoscopically (i.e., 100 – 180 ms, VHF#6), and prevent head movements with a chin rest (VHF#10). However, eye tracking might be more critical if none or only some additional control measures are in place. In addition, the reliability of the VHF technique will always depend on the specific task and stimulus characteristics (Beaumont, 1982). Therefore, our general recommendation is that while many of these statements provide good guidance for setting up VHF tasks, deviations from these recommendations are possible, if researchers explicitly state which parameters they chose and why.

Eye movements. One of the largest agreements was on eye-movement control. Over 92% of the experts agreed that "eye tracking is good but not necessary" in VHF studies (VHF#1). This is in line with Geffen et al. (Geffen et al., 1972) who directly monitored eye movements and reported failures of fixation on only 0.5% of trials. A more recent study (Van der Haegen et al., 2010) came to the same conclusion that strict eye movement control is not needed for valid laterality research. The experts also agreed (77.0 %) that the fixation stimulus should remain visible when parafoveal stimuli are presented to avoid short saccades (<100 ms) (VHF#2). However, the literature suggests that short saccades are not necessarily problematic. Express saccades can be around 100 ms or slightly faster (Fischer & Weber, 1993). They typically occur in trained participants and when there is a time gap between the offset of the fixation cross and the onset of a lateral stimulus (Fischer & Weber, 1993). If there is no time gap (i.e., fixation stimulus disappears at the same time when a stimulus is presented), latencies of saccades are still around 180 to 200 ms (Cohen & Ross, 1977; Fischer & Ramsperger, 1984), and therefore slow enough to allow lateral presentation. Anticipatory saccades can be much shorter than 100 ms and occur when participants can predict when exactly a stimulus will be presented (Fischer & Weber, 1993). Both express and anticipatory saccades can be prevented by varying the duration of the fixation stimulus, by presenting target stimuli randomly to either the left or right side, or by presenting target stimuli to both visual fields simultaneously. This assessment was shared by the commenters who also underlined that eye tracking is not strictly necessary (Van der Haegen et al., 2010) and confirmed the advantage of continuous presence of the fixation cross but also remarked that it does not guarantee absence of saccades. In conclusion, keeping the fixation cross visible during presentation of the lateral stimulus is generally good guidance but not necessarily required for VHF tasks, if other conditions are met.

Task construction. There was strong agreement (89.0 %) that manual stimulus-response compatibility effects should be avoided by switching hands or requiring bimanual responses (VHF#3). It was further suggested by one commenter that response hand could be used as a variable during data analysis. Commenters generally questioned whether counterbalancing was sufficient to solve stimulus-response compatibility effects and suggested that more research on this topic seems warranted. Further, experts agreed (76.9 %) that lateral stimuli should be presented with a duration of 100 to 180 ms (VHF#6). At least two commenters argued that durations shorter than 100 ms would also be acceptable. The statement is roughly in line with the literature which recommended 150 to 180 ms for unilateral presentations (Bourne, 2006), while up to 200 ms are deemed acceptable for bilateral presentations (Hunter & Brysbaert, 2008; Walker & McSorley, 2006). However, commenters noted that stimulus duration depends on the research question.

There was some agreement (65.4 %) that stimulus eccentricities should be 1 degree visual angle off fixation and not exceed 6 degrees visual angle (VHF#8). However, commenters differed in their views whether it should be "at least 1 degree", "arguably greater than 1 degree", or "lower than 1 degree", with the argument that there is no evidence for foveal overlap. The literature recommended minimum values of 2 degrees (Young, 1982) or 2.5 to 3 degrees (Bourne, 2006) when stimuli are presented unilaterally. However, it is noted that eccentricity depends on several factors such as monitor resolution, size of stimuli, complexity of stimuli, and length of stimulus presentation (Bourne, 2006). Thus, while it can be considered good guidance to have at least 1 degree visual angle (possibly better 2 for unilateral presentations), stimulus eccentricity also depends on context factors. The statements about bilateral presentation received little support, in general. Only 38.5 % agreed with the statement that bilateral presentation is better than unilateral presentation (VHF#5). The statement was probably considered too general as many commenters said that presentation type would depend on the research question. The statement "In VHF studies with bilateral presentation and an arrow in the middle pointing to the target stimulus up to 200 ms are possible" (VHF#7) also received only 50.0 % agreement. However, the majority of commenters who did not endorse the statement rated it as either “neutral” (26.9 %) or indicated “no opinion” (19.2 %). Only less than 4 % disagreed. Despite the relatively low consensus rating, however, there is reasonable support in the literature for bilateral presentation (Hunter & Brysbaert, 2008; Walker & McSorley, 2006).

Reliability. There was little agreement about the number of observations per condition (46.2%, VHF#9) and the need to calculate the reliability of VHF differences when analysing individual VHF differences (34.7%, VHF#12). The commenters stated that a minimal number of observations or recommending the use of reliability measures were considered too prescriptive as it depends on the task, the research question, the circumstances, and the number of participants. While most agreed with the advantages of a chin rest, it is also remarked that expected effects are obtained without these requirements which may also cause discomfort when other measures (as discussed earlier) are in place. In an early publication, Satz (Satz, 1977) argued that when there is a strong a priori probability that a certain laterality exists, a laterality test must be highly reliable before it can improve on the null hypothesis that a tested individual conformed to the typical pattern (e.g., an RVF advantage in language lateralisation in a consistently right-handed individual). One critical issue with this is that it is not always clear what the *typical pattern* is as laterality degree (and direction) can vary with task characteristics (e.g., stimulus and task conditions (Smekal et al., 2022)), and with individual factors such as native language, strategies, practice, mood changes, and endocrine factors (e.g., (Hausmann et al., 2002; Hausmann & Gunturkun, 1999; Hausmann et al., 2016).

Analysis pipeline and data reporting. There was good agreement amongst the experts (84.6 - 92.3%) about the statements referring to the analysis and data reporting. That is, laterality indices should analyse response times of only correct responses (VHF#15) *and* accuracy (VHF#14) by taking the overall performance into account (VHF#13). One commenter suggested that it might be worthwhile to explore measures that combine both (Liesefeld & Janczyk, 2019). Also, authors should report LVF and RVF performances in addition to laterality indices (VHF#17). It is indeed good scientific practice to make all data publicly available (e.g., in the appendix or open science platform) and share them with the laterality community.

Outstanding issues

If manual responses are measured, stimulus-response compatibility effects can be avoided. While the interfering nature of compatibility effects is generally acknowledged, it is unclear what measures are most suited to neutralize them. More research on this topic seems warranted.

PERFORMANCE ASYMMETRIES

Recommendations

Defining laterality

- For motor tasks involving objects, clearly define the spatial location of the object relative to each hand (e.g., in terms of a central body point from which deviations are measured; object placed at equal distances from each limb, etc). (PA#1rw, 100% agreement).
- In the Methods Section, explain how it is ensured that there are no implicit and explicit methodological or perceptual biases caused by apparatus positioning. (PA#4, 89% agreement).
- In studies of lateralized interaction with the environment, define peri-personal and extra-personal space. (PA#3, 66% agreement).

Paradigm construction

- Where tasks can be performed from left to right (and starting direction is not part of the research question), counterbalance the starting direction across conditions. (PA#8, 83.3% agreement).
- Decide whether or not a general laterality index of performance should include bimanual activities. (PA#5rw, 72.2% agreement).
- In research that includes bimanual activities in the laterality index, distinguish between symmetrical and asymmetrical activities. (PA#7rw, 72.2% agreement).

Relevant background information

- Acknowledge the possible effects of prior experience on current biases, e.g., keyboard/piano lessons for handedness, soccer training for footedness. (PA#11, 72.2% agreement).

Use of different behavioural measures

- In research that includes bimanual activities in the laterality index, use objective markers to distinguish bilateral activities from unimanual activities. (PA#14rw, 78% agreement).
- For a motor task, assess the direction and degree of lateralization. Therefore, for preference, assess choice of limb, i.e., direction, whether the task was performed by the right or the left limb; for performance assess degree of lateralization, i.e., timing and accuracy measures. (PA#13, 72.2% agreement).
- Performance tests should place reasonable constraints on time and task, and here, as elsewhere, “reasonable” means taking the participant’s age and capacities into account. These constraints should be clearly defined. (PA#16rw, 66.7% agreement).
- Motor performance asymmetry is task specific. Although averaging may sometimes be useful, also report asymmetries for each specific task. (PA#18, 66.6% agreement).
- Tests of motor preferences should reflect validated measures with a known empirical basis and, ideally, should be based on multiple behaviours. (PA#12, 61% agreement).
- If possible, assess both direction and degree of lateralization of lower and upper limbs. (PA#17, 55.5% agreement).

Use of kinematic analysis

- When hand performance is measured through kinematic recording/analysis, attach markers at comparable anatomical landmarks on each hand and across studies. Also explain why these landmarks were chosen. (PA#23, 77.7%).

Errors, validity and reliability

- Quantification of motor performance asymmetry should be based on enough trials to ensure an acceptable degree of retest reliability. Base them on at least two to three trials for each limb. (PA#39, 88.9% agreement; PA#40, 89% agreement; PA#41rw, 66.7% agreement).
- When calculating the LI, clearly explain whether it is calculated from the hand used for reaching (approach phase) or from the hand used for grasping (picking-up phase). In infants, reaching and grasping do not always yield identical LIs. Indeed, it is not unusual for infants to start reaching for an object with one hand, only to grasp it with the other hand. (PA#30, 89% agreement).
- If possible, the task and the dependent variables should be selected to avoid floor and ceiling effects. (PA#44, 88.9% agreement).
- Where possible, videotape performance tests to allow for estimating performance reliability. (PA#38, 61.1% agreement).
- Make the nature of the motor movement explicit and homogenous in order to calculate handedness indices. (PA#43, 55.6% agreement).

Laterality index formulas

- Where a cut-off point is used to categorize behavioural laterality, it should be clearly stated (e.g., consistent footedness is defined as an LI of ± 80 or above; mixed footedness as an LI between -79 and +79). (PA#50, 100% agreement).
- Report the magnitude of the performance asymmetry and the proportion of subjects who show the asymmetry. (PA#47, 66.7% asymmetry).
- Report the percentage performance difference between the preferred and non-preferred body side as it provides a clearer measure of how much the sides are asymmetric. (PA#49, 66.7% agreement).

Comprehensive data reporting

- For assessing asymmetries in manual performance, report mean values as well as variability. (PA#55, 100% agreement).

Performance asymmetry in children

- Base limb preference indices of infants and children on frequency of limb use. (PA#34, 88.9% agreement).
- Parental reports of children's hand and arm preference are not always adequate for assessing laterality. Wherever possible, use performance-based measures of preference, which are more likely to be valid and reliable. (PA#35, 88.8% agreement).
- We will need to agree on a specific set of guidelines that acknowledge the special dynamics associated with performance asymmetry in infants and children. (PA#58, 83.3% agreement).
- Clearly define unimanual versus bimanual reaches and grasps. Strictly speaking, a unimanual reach means that only one hand is used for reaching and grasping; a bimanual reach means that both hands are used. Depending on the child's age, there may be no discernible difference between the hands, either for starting to reach (initiating movement toward the object) or for finishing (grasping the object). Eventually, one hand will take the lead by reaching first and/or by grasping first. Alternatively, the hand that starts second may be the first hand to grasp. A maximum delay between movements must be stipulated for a reach to be considered bimanual. If the second hand does not begin to move until the first hand has grasped the object, this should not be considered a bimanual reach or grasp. (PA#29, 77.8% agreement).
- Calculate an LI using objects that do not afford bimanual manipulation. When objects afford bimanual manipulation, it is more difficult to infer the child's actual intent, i.e., whether the child intended to grasp the object with one hand with the goal of using the other hand for manipulation or to do the reverse. (PA#31, 77.8% agreement).
- Bimanual reaches/grasps are an important part of most infants' behavioural repertoire. Not including bimanual grasps underestimates the number of non-lateralized infants. If researchers decide to remove bimanual reaches/grasps from their LI formula, state the number that were removed. (PA#33, 72.2% agreement).
- In tests with a discontinuous movement parameter (e.g., pegboard tests, dot-filling tests) where milestones of motor development are considered (e.g., adequate difficulty of the test), dot filling and copying line are not recommended for pre-school children; for them, pegboard tests are more appropriate. (PA#28, 61.2%).

Critical review by Lauren J. Harris

We probably can agree that the measurement of motor performance asymmetry pertains to the limb, typically the hand, used to perform, or, for bimanual acts, to take the lead in performing, a broad range of acts. We also can probably agree that to find out, the usual procedure is to observe an individual performing each act while measuring the quality of performance on such features as speed, fluency, accuracy, and force, and, when the same act is repeatedly performed, by its consistency with respect to these same features. Most if not all details on asymmetry measurement were incorporated into 46 statements organised into nine categories.

Defining laterality. Panellists agreed that potential biases of object and apparatus positioning should be acknowledged and disclosed. One panellist cited Sainburg and Schaefer in this connection (Sainburg & Schaefer, 2004), but recognised that these biases may be hard to control, especially when testing children. Panellists also agreed on the value of adopting a context-specific definition of peri-personal and extra-personal space.

Paradigm construction. The majority of panellists favoured deciding whether bimanual activities should be included as part of a general laterality index, but neither option received a majority of votes. Those favouring inclusion mentioned the enhanced ecological validity that can ensue (Boklage, 1980) or at least recommended including them as a separate measure. Those favouring exclusion, while acknowledging the fundamental postural, neural, and developmental differences between bimanual and unimanual actions, noted their possibly poor quality for measuring handedness. Panellists also agreed that counterbalancing the starting direction across conditions is preferable to testing the preferred limb first but noted that this may depend on the research question.

Relevant background information. There was only one statement in this category, and it was agreed to by 72.2% of panellists, namely, that researchers should take note of the possible effects of participants' prior experience when measuring laterality. What was not asked was how it should be considered, e.g., based on amount or length of experience or on skill level attained, and the kinds of experiences to consider (only two examples were given: keyboard/piano experience for handedness; soccer experience for footedness).

Use of different behavioural measures. There was agreement to all 6 statements in this category but only by relatively modest majorities (55.5 – 78%). Some panellists complained about the vagueness of certain statements such as "Taking note of the possible effects of prior experience of current biases," "using validated measures that have known empirical basis," or "performance tests should place reasonable constraints on time and task." They asked, what does "taking note" imply, what "empirical basis" should be required, and what is meant by "reasonable"? Some statements also were criticized for what was seen as their overly prescriptive, and what one panellist called, their "imperative style." For example, while most panellists would recommend assessing direction *and* degree of lateralization, they noted that this might not *always* be the case; nor would it be relevant to assess *all* measures that quantify aspects of performance asymmetry. And while acknowledging the relevance of footedness and other lateral preferences (Marim, 2011; Packheiser et al., 2020), their assessment might not always be relevant for the research question. Similarly, panellists generally agreed that motor performance asymmetry is task-specific and recommended reporting asymmetry per task (Buenaventura Castillo et al., 2020; Packheiser et al., 2020), but they also noted that averaging over tasks and objects is useful for providing a general idea of a person's handedness.

Use of kinematic analysis. Just over a quarter (27.7%) of panellists favoured measuring limb use at the level of kinematics, citing several reports in support (Mathew et al., 2019; Sainburg & Schaefer, 2004; Schaffer et al., 2020). The rest were neutral, had no opinion either way, or disagreed, noting that it isn't always necessary and might be unavailable for many researchers by requiring specific equipment and expertise in kinematic data analysis. A large majority (77.7%), however, agreed that when used, marker position should be clearly reported (Sainburg & Schaefer, 2004) while acknowledging that different systems use different biomechanical models and that consensus on landmark references can be difficult to establish.

Errors, validity, and reliability. Support was mixed for performance measures that combine speed and accuracy. While some panellists noted the possible usefulness of a single index of performance (e.g., (Musalek et al., 2015), others noted that these are not the same and therefore should be reported separately. And while many urge the use of videotaping, others saw disadvantages because of issues pertaining to ethics approval and data storage. Several panellists proposed a minimum number of observations to obtain reliable performance estimates but that the number should be empirically based; they also noted that the number could depend on the type of task and its intertrial reliability. Others raised the issues of practice effects and transfer skill with repeated trials. To avoid the 'number of trials' issue, one panellist proposed a high-enough number to ensure an acceptable degree of retest reliability, but others noted that the vagueness of 'acceptable degree' renders the statement meaningless.

Laterality index formulas. Panellists showed considerable variability in their views about laterality index formulas. There also were criticisms of several statements about specific methodological considerations for its calculation. They were regarded as unclear about their purpose or too prescriptive. A panellist who advised alternative approaches to categorizing was reminded that (established) cut-offs have the advantage of comparability across studies. Some panellists also emphasized the importance of clarity where it comes to the choice and rationale for choosing a particular cut-off point. Ideally, they said, they should be specified in advance, pre-registered, and chosen so as to reflect a consensus judgment among researchers.

Performance asymmetry in children. Statements about measuring performance asymmetry in children were generally well-received (unless they, too, appeared to be overly prescriptive or were unclear, or both) as were those acknowledging reports of assessment differences between children and adults (Campbell et al., 2015; Fagard et al., 2017). Nevertheless, several panellists noted the advantages of maintaining consistent guidelines (and tasks) over the lifespan.

Glancing over the range of votes across statements shows that although the panellists' ratings ranged widely, there was high agreement overall, with on average nearly 67% agreeing and only 10% disagreeing. And of those who agreed, nearly a third (28.3%) agreed strongly, whereas of those who disagreed, only 2.8% disagreed strongly. Of the rest, just over 18% were neutral and only 5% had no opinion.

Most of the statements in the survey could be said to lack an explicit rationale or reason, and the possibility arises that this could be why certain statements lacked agreement. An example is PA#8 in the category, "Paradigm construction" where, if the participant identifies one limb as preferred, the preferred limb should always be tested first. To this statement, 44% disagreed strongly. Had a reason been given, perhaps more would have agreed. Another example is PA#33 in the category "Laterality index formulas." Could agreement have been greater than 38.9% had the justification for the formula been explained?

These possible problems notwithstanding, the overall high level of agreement is reassuring insofar as it shows that most laterality researchers are on the same page when it comes to the measurement of performance, keeping in mind that agreement is greater in some categories than others. For those with strong agreement, there would be reasonable justification for establishing standards for researchers to follow. For those lacking strong agreement, there would be reasonable justification for proceeding with caution.

Outstanding issues

Experts agree that a clear definition of peri-personal and extra-personal space should be provided in studies dealing with these issues. It might be helpful if a consensus could be reached on a general definition of these concepts that is applicable to most standard situations.

When asked whether or not a general laterality index of performance should include bimanual activities, most experts agree. Unfortunately, they do not agree whether bimanual activities should be included or excluded, so this issue remains open. It might be relevant to list in more detail the arguments for or against and come up with a consensus for a ‘general’ laterality index.

Researchers should take note of the possible effects of prior experience on current biases, e.g., keyboard/piano lessons for handedness, soccer training for footedness. While many experts agree with this statement, it remains undetermined how such prior experience should be considered, both conceptually and mathematically.

Where bimanual activities are included in the laterality index, objective markers are needed for distinguishing them from unimanual activities. While not everyone might favour including bimanual activities, it seems advisable to come up with recommendations for distinguishing them from unimanual activities.

Performance tests should place reasonable constraints on time and task, and here, as elsewhere, “reasonable” means taking the age and capacities of the participant into account. These constraints should be clearly defined. The definition of what can be considered reasonable could be empirically validated.

Experts agree that quantification of motor performance asymmetry should be based on enough trials to ensure an acceptable degree of retest reliability. They recommend basing them on at least two to three trials for each limb, but questions are raised as to the empirical support for this recommendation.

We will need to agree on a specific set of recommendations that acknowledge the special dynamics associated with performance asymmetry in infants and children.

PREFERENCE BIAS REPORTS

Recommendations

Setting the standards

- A consensus should be reached determining a gold-standard for assessment of hand preference. (PBR#1; 77.8% agreement). Unfortunately, when presented with different options, no majority was found for either option.
- When relevant, inventory items should take cultural differences into consideration. In cross-cultural settings, inventory items should be limited to tasks that are common across a broad range of cultures and should be translated into multiple languages; culturally-biased behaviours (e.g., feeding, grooming, or social greeting actions), where included, should be flagged for exclusion when inappropriate. (PBR#6rw; 77.8% agreement).
- Clear analysis standards/procedures should be developed and agreed upon to create individual and population-level laterality profiles that include assessments of both sensory (visual, auditory) and motor (handedness, footedness) biases. (PBR#7; 73.3% agreement).
- A complete laterality assessment should include both i) a measure of preference (e.g., Edinburgh Handedness Inventory) and ii) a direct assessment of relative skill (e.g., Pegboard task). (PBR#5; 55.6% agreement).

Definitions

- Fully describe all inventory items; if inventory items are part of an established, validated questionnaire (e.g., Edinburgh Handedness Inventory), authors should provide a reference where a complete description may be found. (PBR#9; 95.5% agreement).
- Classification labels (i.e., weak, moderate, or strong left- or right-handedness; mixed-handedness, ambidexterity) require clear and consistent definitions. Definitions should include scoring criteria (i.e., applicable ranges) from common/standard assessments. (PBR#10; 93.3% agreement).
- A standard glossary of terms should be developed to support meta-analyses and systematic reviews. Terms requiring definition: handedness, footedness, eyedness, earedness, preference (hand/foot/eye/ear), lateral preference, motor laterality degree/degree of laterality, laterality indices. (PBR#8; 88.9% agreement).

Test construction

- Because laterality is a multivariate construct, measure laterality using a multiple-item questionnaire to allow assessment of degree, as well as direction, of preference. (PBR#14; 88.9% agreement).
- Hand Preference/Skill is not universal; one hand may be dominant for a given task (or set of tasks), while the other may be dominant for another task (or set of tasks). Questionnaires therefore should measure preference across multiple criteria. Suggested criteria include fine motor skills (e.g., writing), gross motor skills (e.g., swinging a bat), open/cyclic/continuous actions (e.g., stirring a pot), closed/discrete actions (e.g., reaching, grasping), ballistic actions (e.g., throwing), and communicative gestures (e.g., waving, pointing). (PBR#11; 84.4% agreement).
- We must decide whether or not a general laterality index of preference must include bimanual activities. (PBR#16rw; 80% agreement). Unfortunately, when experts were presented with the different options, no majority was found for either option.
- Inventory items should have a fixed number of response options. (PBR#12; 71.1% agreement). A small majority (53%) is in favour of a 5-response options solution: always left, mostly left, either/no preference, mostly right, always right.

Laterality index formulas

- If the test includes bimanual items, each item should define the hand used for the active part of the action (for instance for scissors, the hand holding the scissors for cutting should be the one reported, not the one holding the sheet of paper, etc.). (PBR#19; 93.3% agreement).
- Consensus is needed on how to handle ‘both/either’ in quantifying hand preference. (PBR#17; 75.5% agreement). If both/either responses are used, it is recommended to add the ‘either hand’ options to the Lis denominator in a laterality index, such that non-lateralized actions lower the calculated LI (PBR#17bis; 64.4% agreement).
- If bimanual activities are probed in the laterality index, distinguish between symmetrical and asymmetrical activities. (PBR #18; 64.4% agreement).
- Design response paradigms to enable respondents to give one simple answer per question. (PBR#20; 62.2% agreement).

Reliability and validity

- Provide the psychometric validation of an inventory or questionnaire before using it for a specific purpose. Any variations/modifications should be clearly labelled and described. (PBR#23; 82.2%).

Infants and young children

- Self-report and parental report via survey are not sufficient indicators of manual biases – especially in the case of children. With younger children, a test with real objects (‘show me how you usually use this pencil, etc.') is better than a pretend execution of the action without objects ('show me with your hands how you use a pencil, etc') and even better than a questionnaire ('with which hand do you?'). (PBR#24; 82.2% agreement).

Scoring and classification

- Ambidexterity should refer to equality of performance, rather than ambiguity in preference. In other words, ambidexterity denotes an equality of performance ability, regardless of (daily) preference. If one writes equally well with both hands, but prefers one hand most of the time, then this person is ambidextrous. Preference reports should make this distinction. (PBR#29; 64.5% agreement).
- The term ambidextrous should be used to indicate no preference between left and right for one specific task (e.g., individual can write with both left and right hand) and the term mixed handedness should be used to indicate preference of different hands for different tasks (e.g., use the right hand for some tasks and the left for others). In the case of laterality indices, we should avoid using the term ambidextrous and mixed-handedness and rather refer to scores in a range around zero representing equal performance between the two hands on a specific task. (PBR#28; 60% agreement).
- Consider multicollinearity of used items in measures of hand preference. Probing the hand preference for 7-8 tool items, still probes only the 'tool' construct of hand preference, while other factors of laterality preference remain unexplored. (PBR#30; 55.1% agreement).

Critical review by Jason Flindall

Setting the standards. While the majority of respondents agreed on the need for a 'gold standard' evaluation tool for evaluating hand preference (more consistency, better comparison between studies), there is yet no existing tool that fills the necessary requirements. Respondents to this survey posited that an ideal gold standard would be:

- simple to understand, appropriate for a broad range of ages and education levels;
- brief (quick to administer, appropriate for use in large studies);
- comprehensive; i.e., able to measure multiple lateralized behaviours, including degree and direction of asymmetries in manual preference and skill (see PBR#11 and PBR#14);
- able to quantify handedness both categorically (to support statistical analyses, cross-study comparisons and meta-analyses) and continuously (to reflect the variability and inconsistency of the construct; see PBR#3);
- culturally unbiased, appropriate for the broadest possible range of participants (PBR#6); and
- cross-validated, such that it can be used to predict/infer asymmetries in sensory, neural, and motor domains (PBR#7).

Some of these requirements were in direct contradiction with each other, making the establishment of a single gold standard difficult. For example, the requirements to be comprehensive would require the gold standard tool to be able to assess the degree and direction (see PBR#14) of asymmetries in: fine motor actions (e.g., writing); gross-motor actions (e.g., swing a bat); open/cyclic actions (e.g., stirring a pot); closed/discrete actions (e.g., reach-to-grasp, or reach-to-point); ballistic actions (e.g., throwing); communicative gestures (e.g., pointing, waving); asymmetric bimanual actions (see PBR#16bis); and tool-use. It should have components for assessing both preference, via self-report, and skill, via demonstrative action (PBR#5, PBR#11). It should be simple and unambiguous, with a fixed number of options per question from which participants may select their responses (PBR#12, PBR#13). Needless to say, no currently existing tool met these requirements, let alone one that is also brief.

Among existing tests (Marim, 2011; Oldfield, 1971; Steenhuis & Bryden, 1989), the Edinburgh Handedness Inventory (EHI) (Oldfield, 1971) is the most popular candidate, selected by just under half of those respondents in favour of adopting a gold standard. While the EHI has the benefit of being validated, well-known, and brief, it both lacks a demonstrative component and contains items ("With which hand would you strike a match?") that made it inappropriate for assessing hand preference among children. It also includes psychometrically invalid items ("With which hand would you lift a lid?") and employs an arcane scoring-system that some participants may find confusing. Modified versions of the EHI have been developed to address some of these shortcomings, but none of these have yet found universal support among laterality researchers. Several commenters questioned the need for such a gold standard as it should be adaptable to contemporary habits (all sorts of electronic devices) while others provide references to recently adapted versions of questionnaires (Dragovic, 2004; Edlin et al., 2015; Fazio et al., 2012; Leppanen et al., 2019; Lyle et al., 2008; Milenkovic & Dragovic, 2013; Prichard et al., 2013; Prichard et al., 2020; Prieur et al., 2017). Further comments ranged from a disbelief in preference questionnaires (a vast difference from performance measures) to the suggestion of creating a new inventory. Regardless of which tool one uses, respondents almost universally agreed that all inventory items used to assess handedness should be fully-described and, where possible, linked to their validation studies (PBR#9).

Definitions. A gold standard assessment tool would facilitate meta-analyses and systematic reviews. Along the same lines, there is a pressing need to standardize commonly-used terminology in laterality research. When directly asked, 88.9% of respondents agreed that a glossary should be developed to define ambiguous nouns like *handedness* and *footedness,* and potentially-subjective adjectives like *strongly* or *weakly*. 93.3% of respondents agreed that these adjectives, if used to describe one's degree of preference, should be accompanied by clear and consistent definitions and scoring criteria.

Test construction. With regard to test construction, another key term is *bimanual*. While 80% of respondents agreed that it is important that we come to a consensus on whether bimanual actions *should* be included in the calculation of a general laterality index (PBR#16rw), the same respondents can be split nearly equally into groups *for* inclusion (43%) and *against* it (57%). Some of this disparity stems from variable definitions of *bimanual*, with some respondents opting for exclusion until a consensus definition can be reached. Two types of comments emerged. The first focus on the concept of bimanual actions; its specificity and relative importance for a general laterality index. The idea here is not to include them, or only when relevant for the research question. The second type of comment advocated including bimanual actions as they increase the validity of a multivariate phenomenon. Regardless, a strong majority (93.3%) agreed that *if* they are included in assessments, scoring on asymmetric bimanual activities (e.g., opening a jar, using scissors) should differentiate between the manipulating hand and the hand providing support and stability. Ultimately, bimanual tasks may need to be separated from unimanual tasks for assessment purposes (as lateralization for one may not always predict lateralization for the other).

Laterality index formulas. Handling 'both/either' responses in quantifying hand preference and defining the active part in bimanual actions are issues that affect a laterality index. Some of these recommendations are able to convince many experts, others are criticized for being difficult to understand. The demand for psychometric validation of preference inventories is underlined and some proposed large-scale reliability/validity studies of current measures that could be cited where appropriate.

Classification issues. Finally, confusion about the inconsistent use of terms like *ambidextrous* and *mixed-handedness* in respondents' comments on this survey underscored the need for a glossary of terms. Sixty percent of respondents indicated they agreed with the statement that "ambidexterity should be used to indicate no preference between left and right for one specific task"; in a separate response, 64.5% indicated their agreement that "ambidexterity should refer to equality of performance, rather than ambiguity in preference." In spite of their apparent contradiction, each of these statements about ambidexterity had majority support, demonstrating the word's mercurial definition in the minds of many researchers. *Mixed-handedness* had comparable ambiguity, being used to describe either inconsistent lateralization across a variety of tasks, or else interchangeably with *ambidextrous*. Until consensus definitions for these terms achieve widespread adoption, researchers must be careful to clearly and explicitly define their intended meanings when using them in their publications.

Outstanding issues

A standard glossary of terms referring to manifestations of lateralized preference should be developed to support meta-analyses and systematic reviews.

Several statements note that handedness is a multifactorial trait, which, in the same individual, could differ for a given task (or set of tasks). While many experts might agree, there is considerable debate about the number and type of components/criteria/factors that would constitute such a multifactorial approach. A summary of the available empirical evidence and some general recommendations might be warranted. This might also be a starting point to design a reference handedness questionnaire that many experts would be willing to use. At this time, such a questionnaire does not seem to be available as the ones that are available do not seem able to convince a majority of experts.

There is considerable debate on whether to include bimanual actions in a laterality index. Some favour including bimanual actions to increase ecological validity; others feel that defining bimanual actions is tricky and that including them might only add noise to an estimation of laterality.

Several statements on ambidexterity were submitted that were agreeable to most experts, although comments were critical about the fitness of the term and its conditions of use. It could be helpful to come up with some recommendations on its definition in laterality research, on its distinction from related concepts (such as mixed handedness), and the conditions under which it can be used appropriately (for example only in performance context).

ELECTROPHYSIOLOGICAL RECORDING

Recommendations

Recording standards

- The EEG signal should be of comparable quality (e.g., SNR, impedance) across the two hemispheres. (EPR#2; 100% agreement).
- Ensure that the recorded activity is not a result of lateral eye movements. (EPR#3; 100% agreement).
- Arrange the experimental setup to be as symmetrical as possible with respect to the participant's midline. (EPR#1; 94.1% agreement).

Reference schemes

- The choice for a specific EEG reference needs to be clearly stated and justified, and its implications for data laterality analysis need to be reported/discussed. (EPR#15; 94.1% agreement).
- In meta-analyses of EEG asymmetry, include the reference montage as a factor that might explain the effect size. (EPR#17; 88.2% agreement).
- Base EEG asymmetry indices on data that have been current-source density transformed to provide more precise estimates of local laterality. (EPR#4; 58.8% agreement).
- Preferably, reference-free EEG analysis methods at sensor and brain level should be considered and used. CSD or other reference-free measures have some advantages, but may not always be appropriate, particularly with low-density arrays. (EPR#5; 58.8% agreement).

Calculating and reporting asymmetries

- Lateralization of EEG activity should be computed between homologous (groups of) pairs of electrodes across the two hemispheres. (EPR#7; 100% agreement).
- EEG asymmetries should not be phrased as 'higher' or 'lower' symmetry, but the phrasing should always include the direction of asymmetry (e.g., greater relative right frontal activity, greater relative left hemisphere activation.). (EPRs#14, 100% agreement).
- For studies of individual differences in EEG laterality, report the reliability of the laterality index (e.g., split-half or test-retest). (EPRs#18; 76.4% agreement).
- Frontal EEG asymmetry should be reported as ln(right) – ln(left) alpha activity (8 -13 Hz), with higher scores putatively indexing relatively greater left frontal activity, and lower scores indexing relatively less left frontal activity. (EPR#8rw; 64.7% agreement).
- It is also possible to use the laterality coefficient (LC) computed as LC = (R-L)/(R+L), where R denotes alpha power at the right hemispheric electrode position and L denotes alpha power at the homologous left hemisphere position. For mathematical reasons, in the small physiologically expectable range of relative differences between the EEG alpha power at two homologous electrodes, the correlation between LC and the metric (lnR – lnL) is very close to 1. Compared to the metric (lnR – lnL), the range of LC is confined to -1 to +1, which makes the meaning of scores intuitive; and as LC is a relative score, hemispheric differences are easily comparable between different electrode positions, conditions, and studies. LC is also commonly used in other fields of laterality research. (EPRs#9, 64.7% agreement).

Data analysis and reporting

- Along with reporting laterality indexes (or laterality tests in MANOVA), report effects in each hemisphere. (EPR#11; 82.3%).
- Reports of significant correlations with laterality indices should be followed with correlations on each side. (EPR#12; 82.3% agreement).
- If laterality (effects) is (are) predicted in specific areas, compare them to laterality (effects) in other control areas. (EP #13; 70.5% agreement).

<u>Critical review by Nicolas Badcock and Gina Grimshaw</u>

A number of publications have specifically addressed methods for recording, describing, and analysing EEG asymmetries (Allen et al., 2004; Smith et al., 2017), and the generated statements broadly cover the same issues. These statements fall into three categories.

Recording standards & Reference schemes. Commenters underlined the general agreement toward a symmetrical setup (Schneider et al., 2012), comparable bilateral signal quality, and controlling for the effect of eye movements (although this might be hard to achieve). There was less agreement about reference schemes. About half of respondents advocated the use of Current Source Density (CSD) transformation or other reference-free measures (Burle et al., 2015; Kayser & Tenke, 2015b; Tenke & Kayser, 2005, 2012), but commenters noted that these measures may not be appropriate with low density recording arrays (i.e., < 60 channels) and would not support it as a general principle. While comments supported data sharing, there was low endorsement for using the standardised Brain Imaging Data Structure (BIDS-EEG) to facilitate public sharing of data, but these standards are relatively new, first appearing in 2019 (Pernet et al., 2019).

Calculating and reporting asymmetries. The proposal for a minimum number of 100 artefact-free epochs was found to be arbitrary; while a large number of trials is preferable, it was remarked that this is not the only variable to consider (Cohen, 2015; Kayser & Tenke, 2015a) and might not be feasible in all experimental designs. The number of epochs contributing to each variable should be reported.

Commenters also criticised the focus on frontal alpha activity over other frequencies and locations (Allen et al., 2004; Tenke & Kayser, 2005). The literature on asymmetries in electrophysiological (EEG) measures is dominated by research on asymmetries in alpha power (8-13 Hz) recorded over frontal sites (Smith et al., 2017). But asymmetries can be measured in any frequency band and over any pair of (usually) homologous electrodes (Ocklenburg et al., 2019). Despite unanimous agreement for homologous electrode comparison, one expert remarked that exceptions can be made if there is a convincing rationale to use non-homologous scalp recording sites (e.g., that consider the anatomical differences between the two hemispheres).

Asymmetries can also be calculated in event-related measures (ERPs), and so any means of measuring and reporting laterality should apply to the range of possible measures. By convention, these asymmetries are calculated by subtracting left from right measures, usually of EEG power. Two indices were proposed: lnI – ln(L) is commonly reported in the frontal asymmetry literature, but others advocated for a normalised (R-L)/(R+L) index, which has the advantage of being bounded by +1 and -1 and being analogous to many laterality indices calculated in other domains (e.g., in handedness, or perceptual asymmetries). One commenter correctly noted, though, that in the range of asymmetries that are anatomically possible (at least in people with two hemispheres), the correlation between these two indices approaches *r* = .99 (Allen et al., 2004). It may not matter, therefore, which index is used. What is most important, therefore, is that left and right hemisphere values are reported alongside any LI, and that raw data are shared so that alternative indices can always be calculated. This will be especially important for meta-analyses.

An important concern (and source of confusion) in this literature is the difference between EEG power (which can be measured) and cortical activity (which is inferred). This confusion arose because cortical "activity" is thought to be inversely related to alpha power. Thus, an index that describes greater right than left alpha power is interpreted to show greater left than right activity. This may be true only of frontal asymmetry and only in the alpha band; given the wide range of neurocognitive functions indexed by EEG oscillations in different bands, this cannot be assumed to be universally true. For this reason, there is agreement that researchers should always be very clear about what asymmetry they are describing (e.g., observed power, or inferred activity), and describe asymmetries in terms of relative attributes (e.g., greater left than right parietal activity, greater right than left frontal alpha power) so there is no ambiguity as to the asymmetry reported.

Data analysis and reporting. The final category of statements concerned analytical approaches. Here there was little agreement, with several researchers commenting that analytical approach depends on the research question to be asked and the nature of the data available and cannot be generalized (Kayser & Tenke, 2003, 2005). Some advocated for data-driven approaches, while others thought pre-registered, hypothesis-driven approaches are preferable. This range of opinions highlights the need for clear reporting of data (e.g., of individual hemisphere values as well as any laterality index) and data sharing to allow alternative analytic strategies to be applied. While clear reporting was agreeable to many experts (reporting reliabilities, reporting values for each hemisphere separately), several commenters also noted that necessary reporting depends on the research question and specific measures should not be mandatory.

FUNCTIONAL TASK-RELATED MRI

Recommendations

Tasks and paradigms

- When assessing language dominance, use several tasks and express lateralization in terms of LI for language subcomponents (production, comprehension, semantics, phonolI...). (fMRI#4, 82.6% agreement).

Reliability

- When calculating laterality indices, routinely assess the quality of the underlying data. (fMRI #9, 91.3% agreement).
- In clinical care, use laterality index-based predictions of neurosurgical risk to cognition only if the protocol they are based on is precisely replicated. (fMRI#10, 78.6% agreement).

Region-of-interest

- Make publicly available the regions of interest and reference brain used to calculate laterality indices with documentation on their source and construction. (fMRI#16, 91.3% agreement).
- When calculating the laterality index, take into account the region's size or to use regions with comparable sizes (fMRI#15, 86.9% agreement; fMRI#22, 65.2% agreement).
- All voxel-based measures of lateralization should be performed only after spatial normalization to a symmetric template. (fMRI#11, 69.5% agreement).

Analysis (Method)

- Basic mathematical operations (addition, subtraction or multiplication) on LI's are not recommended. (fMRI#31, 56.5% agreement).

Analysis (thresholding)

- When selecting a given method to calculate (threshold-free) laterality indices, provide a clear rationale why this method was chosen. (fMRI#35, 95.7% agreement).
- When calculating an LI, use methods that use multiple significance thresholds or that are threshold-free (fMRI#33, 73.9% agreement; fMRI#36, 69.5% agreement; fMR#32, 52.2% agreement).

Reporting

- Motivate and report the choice of the contrast to assess functional laterality of a given cognitive function. (fMRI #37, 95.6% agreement).
- Clearly define 'left', 'right' and 'mixed/bilateral' (language) dominance. (fMRI#46, 78.3% agreement).
- In clinical care, determining language dominance should refer to a LI, but should not be based only on that LI. (fMRI#47, 65.2 % agreement).

Critical review by Marc Joliot and Karsten Specht

Overall, there was agreement with the statements with an average of 93% (range 70-100) of positive responses (sum of « agreed » answers normalized to the sum of responses not taking into account the neutral and no opinion answers). The added neutral and no-opinion votes were around 25% [range 4-61] for each question, with at least half the panel having a positive or negative opinion, except for one question (#27). Overall, this precluded having answers dictated by only a few individuals. Nevertheless, the very low "disagree" percentage (average of 5%, range 0-23) suggests that there was a bias in how we designed the questionnaire, which possibly led to the inflation of agreement.

Tasks and paradigms. There was general agreement that more research is needed to find the optimal task(s) (and control(s)) to compute the LI. However, if it is desirable to compare different studies, it is our belief (as pointed out by some of the comments) that the goal will be tough to achieve. In particular, the use of "rest" as a control condition is part of a controversial discussion. It also is possible but not desirable as it will create a bias toward one component of the studied function. For language, for example, it has been shown that the laterality of production, reading, and listening are not the same in each individual (Labache et al., 2020). Therefore, it would be advisable to not use only one paradigm, even if only speech dominance is assessed, since the estimated or observed dominance might depend on factors related to the paradigm or stimulus material used. The comments acknowledged the usefulness of having several tasks to determine language dominance but also noted that this may depend on the particular research questions of the study, for example, an LI derived from a single speech production task may suffice if speech dominance is the topic of interest. While most experts agreed that a hand preference fMRI paradigm should be established, some commenters argued this would be of limited value.

Reliability. For the "reliability" issues, while there was good agreement about the five questions as in other sections, there was less agreement about the definition of a bilateral representation (see also previous comment and the "reporting" section). Regarding the statement that participants cannot be reproducibly categorized as bilateral, commenters were in favour of reaching a consensus on how to define bilaterality, with one remarking that we currently do not know how reproducible bilaterality is. The importance of establishing test-retest reliability of LI's for standard fMRI paradigms is acknowledged, although it is noted that this may depend on methodological decisions such as the LI calculation method, ROI selection, and population of interest. (Matsuo et al., 2021; Otzenberger et al., 2005; Rutten et al., 2002). There was near unanimous agreement with the statement that the quality of the data the LIs are calculated from should be assessed to avoid issues of data scarcity, with a commenter adding that data quality should be considered as well.

Regions-of-interest. Normalization to a symmetrical template for voxel-based analyses was well-supported. According to one of the commenters, this may be particularly useful in group comparisons (e.g., between left- and right-handers) and another recommended the use of robust normalization methods since it may be challenging to achieve an exact one-to-one correspondence between left and right voxels. There was high agreement on the statement that functional asymmetries can be determined on an ROI or voxel/vertex level, although a caveat was added that this assumes a sensible left-right homology (e.g., based on similarity of intrinsic functional connectivity profiles), such as in (Joliot et al., 2015) and, for a voxel level analyses, a near perfect alignment to a symmetrical template. When the LI is based on voxel count in homologous regions, commenters agreed the regions' size must be taken into account, with one stating that this is necessary to avoid any potential bias. While most statements in this section showed high levels of agreement, statements fMRI#17 and fMRI#20 did not. Both highlight an ongoing debate about how we include the homotopic ipsilateral region if it is not activated (competent) or even deactivated by the studied task. Commenters generally disagreed with the statement that LI's calculated over geometrically homologous regions do not inform about the true lateralization of a function, if one of these regions has no role in said function. Instead, they argued that if a homologous region is not involved in that function, this in fact indicates (strong) hemispheric specialization, in our view, an entirely valid point. One panellist also remarked that a lack of involvement of the homologue does not matter if the LI is clinically relevant, e.g., if it predicts post-surgical recovery. Some comments questioned the statement that ROIs should be small when signal magnitude methods are used. One commenter stated that small ROIs can be valid and that the issue instead is about how signal magnitude should be summarized within a ROI, and another stressed the importance of providing information about the robustness of the measure. Commenters expressed mixed opinions on the statement on the need to establish consensus on how to determine the ROI for calculating an LI for standard fMRI paradigms. One proponent argued this would facilitate comparability of findings between labs, while another who was in favour to some extent suggested the development of a core set of atlases tailored to laterality research accompanied by a set of procedures to translate results between atlases. One opponent remarked that using a single solution will introduce a strong bias, and another reiterated a point made in comments on previous statements that the optimal ROI or parcellation depends on the research question and that this choice should be clearly justified. Regarding the use of ROIs with similar sizes, commenters noticed that, alternatively, size differences can be taken into account by the analysis (e.g., using weighing factors), and another suggested this statement applies only when ROIs are defined anatomically but not functionally. Experts generally agreed it is not adequate to claim activity is lateralized by observing a cluster in one hemisphere without testing whether it is significantly stronger than in homologous voxels in the other hemisphere. However, one commenter mentioned this suffices to conclude that activation is lateralized, but comparison with homologous voxels or ROIs is necessary to make conclusions about the strength of lateralization.

Analysis (Method). This section showed the highest numb“r of "neutral” and “no opinion” votes, which is partly due to the technical nature of the questions. Commenters called into question that measures based on signal magnitude are by default more reproducible and less susceptible to noise compared with those based on signal extent. Whether one outperforms the other may depend on other factors, such as which comparisons are used (Chlebus et al., 2007) or the size of the ROI. Some commenters agreed with the sentiment of the suggestion to calculate an average of LIs based on voxel count and signal magnitude, but favoured a different implementation of combining different LIs into a single metric, for example a metric that takes into account voxel count at different amplitudes. In a similar vein, some advocated reporting LI’s based on both as a way to evaluate the robustness of the findings. The idea that (a)typical laterality needs to be clearly defined is embraced, but it is acknowledged that this will likely depend on the function and that more research is needed to make this possible. Comments that disagreed with the statement that non-parametric test are by default more appropriate than parametric ones to compare LI’s, suggested instead that this decision should be based on the particular distribution of the obtained LI’s. We conclude that overall, this section provides one principal recommendation, namely using both the signal extent and magnitude to define the laterality index. As stated through the comments, those two measures tend to lead to similar (but not identical) measures. From our point of view, it could be done when computing hemispheric index, but in the case of a region of interest analysis, the extent-based index may become challenging to apply or even useless in the case of a small region.

Analysis (Threshold). On the statement advocating the use of unthresholded t-maps or z-maps to calculate LIs in ROIs, one commenter warned that the LI will depend on the chosen percentage and suggested taking all positive voxel values (i.e., to exclude deactivation) while recognizing that this may result in lower LI values on average. Similar to previously made comments in this section, another remark was that the choice to use unthresholded maps depend on the method used to calculate the LI. Those commenters that supported the suggestion to calculate LIs using multiple thresholds of significance, mentioned that this is a way to test the robustness of the results and to guard against any accidental biases. To summarize, this section points out that we should use an independent threshold method, that is, as stated by one comme“ter, "a remedy for any accidental b”ases."

Reporting. The importance of reaching consensus on how to refer to mixed/bilateral language dominance is underlined by the commenters, with one stating that lack of agreement on a definition is a major cause of confusion in the literature. Another proposed to distinguish between the terms "bilateral" (for individual/patient level and whole hemisphere LI) and "mixed" (for group/population level and multi-region LI). Comments on the follow-up question, i.e., which term would be favoured, included various suggestions for names not included as an answer, such as "no hemispheric dominance", "bilateral representation", "bilateral" (i.e., without "dominance") or to use "bilateral" and "mixed" to refer to different concepts. One commenter preferred "mixed" over "bilateral": whereas bilateral suggests equal contribution of the two hemispheres, "mixed" makes it clear that both are involved even if their relative contributions may be complex. Another opinion expressed here is that a clear specification of the range of mixed/bilateral representations and how it was defined (e.g., cut-off, threshold, ROI choice, LI calculion...) is more important than reaching a consensus on its terminology. In general, commenters rejected the proposal that LIs should not be compared between tasks with very different baselines, arguing that this is not an issue since the LI is a relative measure and that depending on the tasks, baselines have to be different (e.g., when they use different input modalities). One commenter, who agreed that a threshold of 0.2 is more objective to generate categorical indices when three categories are included, noted that the interpretation of 0.2 depends on the LI computation method and that this is particularly recommended for threshold-dependent methods based on voxel count (Seghier, 2019). Commenters agreed that in clinical care, determining language dominance should be based only on an LI, with one stating that the effects of brain damage/tumour on brain activation, functional reorganization during recovery, and cases with mixed laterizations warrants the need to refer to the original fMRI maps in addition to the LI (Seghier et al., 2011). The principal advice, in our view, that follows from this section is that each step of the procedure and not only the final categorization should be reported. As evident from the discussions in the other sections, any LI is dependent on various factors, like the task, the type of control condition, input to the LI estimation, and the way LI is estimated. Therefore, it must be emphasized once more that every step and decision need to be documented and justified to achieve a reliable and interpretable LI. As in the second section on the question of the bilaterality, if we agree that we should reach a consensus on the naming, the pool shows (question #40) that there is, in fact, no consensus.

Outstanding issues

Reaching consensus on how to refer to 'mixed/bilateral' language dominance would be beneficial (fMRI #39). However, experts do not agree on what the preferred term should be (Task-related fMRI #40).

Although experts recommend threshold-independent methods to calculate LI (Task-related fMRI #33 and #36) and to motivate the choice for a particular LI calculation method (Task-related fMRI #35), it remains unresolved whether signal magnitude-based approaches are more reproducible and less affected by noise compared to signal extend-based approaches (Task-related fMRI #25). The field may benefit from more methodological studies that compare the performance of both approaches, evaluate how the performance may interact with other methodological decisions (e.g.: ROI size, see Task-related fMRI #20), and test the possible advantages of combining both approaches (Task-related fMRI #26 and #27).

Agreeing on a definition of 'atypical' laterality was deemed to be useful (Task-related fMRI #29). Achieving this will likely prove difficult and involves first dealing with several other issues, including 1) obtaining a better picture of the influence of LI computation methods on the appropriate threshold to generate categorical indices (Task-related fMRI #45) and 2) and collecting more empirical data that could guide the definition process, for example, data on the reproducibility of 'bilateral/mixed' categories (cf. Task-related fMRI #5 ratings and comments).

Experts supported several statements calling for standardization, including for the determination of regions of interest for standard task-related MRI paradigms (Task-related fMRI #21), the selection of optimal control conditions (Task-related fMRI #2), and the development of a hand preference fMRI paradigm (Task-related fMRI #3). In line with these statements, experts are strongly in favour of establishing test-retest reliability of LI's for a number of standard task-related fMRI paradigms (Task-related fMRI #6). All this leaves room for future work.

STRUCTURAL MRI

Recommendations

Spatial normalization and brain (a)symmetry

- Given the asymmetric nature of most brain templates, laterality research should register subjects on a symmetric template to estimate structural asymmetries (SMRI#1, 80% agreement; SMRI#5, 75% agreement; SMRI#7, 85% agreement). Another solution is to stay in subject's space (SMRI#2, 70% agreement).
- To take individual differences in landmark patterns into account, identify the laterality of particular anatomical metrics manually, design new tools for measuring these individual landmarks, or use probability atlases for areas known to be highly individual (SMRI#3, 65% agreement).
- Create a sample-based symmetrical template as it more strongly respects individual differences of the sample (SMRI#4, 65% agreement).

Comprehensive measurement

- For the measurement of brain structural asymmetries, take into account the dissociation between hemispheric differences in sulci position and tissue compartment density or volume hemispheric difference. (SMRI#7, 75% agreement).
- To avoid cherry-picking significant results, first establish agreement on which dependent variables (FA, D, cortical thickness, surface area, etc.) to use to calculate laterality indices (SMRI#9, 70% agreement). However, no consensus is reached on which ones to choose. To avoid cherry picking, the experts recommend preregistration and justification of the chosen variables.
- In case of manual ROI definition, it is important to avoid biases in ROI delineation. Therefore, examine inter-operator reliability to make sure the operator is blind to which hemisphere they are delineating (SMRI#12, 70% agreement).
- For calculating a region-based laterality index, define functional/structural regions for each hemisphere separately based on local anatomical/functional properties. (SMRI#13, 70% agreement).
- For volumetric laterality indices, it is recommended to 1) adjust for brain size (SMRI#17rw, 60% agreement) and 2) adjust for the overall size difference between the two hemispheres (SMRI#18rw, 65% agreement).

Critical review by Karsten Specht

The section on structural MRI consists of 16 statements. Only a few, however, showed a clear consensus, while, on average, the overall agreement rate was only 65% (range [40-85]) when "strongly agree" and "agree" are combined. The following review of the consensus statements is separated into two sections, and each section presents the statements in order from highest to lowest in agreement.

Spatial normalization and brain (a)symmetry. The first section contains five statements about the spatial normalisation of MRI data, especially in connection with voxel-based morphometry studies. There was a clear consensus that symmetric templates should be used for those approaches and refer to recent studies applying such an approach in diffusion MRI (Verhelst et al., 2021) and in surface area and cortical thickness (Xiang et al., 2020). As emphasised by comments, the templates that are implemented in standard neuroimaging software are typically not symmetric. However, it was also pointed out that problems might occur when symmetric templates are created for brain structures where the left and right patterns are not compatible. This might reduce the overall quality of spatial normalisation in this area. Related to this, most researchers also agreed that asymmetrical atlas definitions might help in studies where the goal is measuring individual differences. Although it was still the majority, a weaker consensus was reached concerning the statement that sample-based symmetrical templates should be created. Some researchers expressed the opinion that this is only useful for samples that are very different from the healthy population and when the sample is large enough. However, it is not clear what the critical size might be. It was further agreed that laterality indices are influenced by the used template. Therefore, one might critically ask whether this could also affect the reliability and comparability of studies. One way to circumvent template-based inconsistencies is to explore size differences in the subjects' native space. Differences in landmark patterns in particular might require manual identifications, as most of the panellists agreed. But it was emphasised that manually identifying landmarks is not always easy.

Comprehensive measurement. The second section, with eleven statements, refers to various measures that can be extracted from structural data. Among them are also statements that did not reach a clear consensus and might need further discussion in the community. However, there was an almost complete consensus that anatomically defined regions of interest are an appropriate way to evaluate structural asymmetry. For such an approach, however, regions must be defined properly and coherently. There was also a consensus that varying hemispherical positions of sulci and other properties must be considered in all measurements of brain structural asymmetries. Therefore, all regions should be defined for each hemisphere separately. A consensus was also reached that regions of interest should be defined in a blinded way, such that they are created in a non-biased manner and independent of the operator. A between-operator reliability measurement could evaluate the latter. There also was a reasonable consensus that a set of variables should always be reported, not only those which showed a significant effect.

There was agreement on the statements concerning the adjustment for brain size in statistical measures. However, as several researchers pointed out, this depends on how the laterality index is computed and how the brain size is extracted. Some software tools might provide only the total intracranial volume, which only, to a certain degree, correlates with brain size. Further, some laterality indices might implicitly adjust for brain size. In essence, the consensus was only moderate as the issue is method-dependent.

Only 50% of the panellists agreed that standard MRI sequences, such as T1 or T2, are appropriate only for analysing morphological asymmetries but that microstructural asymmetries need quantitative measurements, such as T1 relaxation times. The background for this statement is that non-quantitative MRI sequences might show systematic spatial inhomogeneities that could influence and hence bias asymmetry measures. It was proposed that those signal inhomogeneities could be accounted for in the processing of the data.

Also, only 50% of the researchers agreed that we need a common approach to report longitudinal changes in asymmetry.

Interestingly, the panellists were more critical of statements related to surface-based methods. Only 45% agreed, 20% were neutral while 30% had no opinion that a laterality index can be computed across the whole cerebral surface on a vertex-by-vertex basis, given an appropriate high-dimensional non-rigid surface registration and the mid-sagittal plane as the anchor. As a potential problem, the use of the mid-sagittal plane was mentioned, as its estimation needs to be robust to the petalia. Further, only 40% agreed while 30% disagreed that surface-based methods should be privileged over voxel-based morphometry because these methods allow a dissociation between cortical surface and cortical thickness asymmetries. However, concerns were raised about the fact that surface-based methods are model dependent and include substantial data processing, which might introduce a yet unknown bias.

Further, sub-cortical structures are not within reach of these methods. On the other hand, surface-based methods might provide specific information that all voxel-based techniques cannot extract. It seems that this method needs further exploration before a clear consensus can be reached.

Outstanding issues

Various statements about structural MRI concern the use of asymmetric templates for studying structural asymmetries (Structural MRI #1, 5 & 7). Most experts agree that symmetric templates should be used, although some outstanding concerns remain. For example, what are the consequences for spatial normalization when averaging highly asymmetric regions to achieve a symmetric template? Is staying in the participant's space a good solution? (Structural MRI #2).

For calculating volume/size-based laterality indices, most experts agree on the need to adjust for brain size (Structural MRI #17rw) and/or for the overall size difference between the hemispheres (Structural MRI #18rw). They do not agree, however, on how to achieve this. It therefore seems important to investigate the possible effects of brain volume on laterality indices and to decide on how to take it into account.

FUNCTIONAL TRANSCRANIAL DOPPLER ULTRASONOGRAPHY

Recommendations

Study set-up

- Assess the participant's behaviour during the POI when possible, which can be achieved by favouring overt or active tasks. (fTCD#5, 92.3%).
- Present different trials randomly where possible (fTCD#2, 76.9% agreement).
- Assess motor activity during the fTCD task so that asymmetrical motor activity, which is a potential confound, can be reported. (fTCD#7rw, 69.3% agreement).
- Include at least 16 trials as the LI is likely to be unreliable if fewer are used. This also implies excluding participants with fewer than 16 trials per condition. (fTCD#1rw, 69.2%).
- Avoid comparing activity during the period-of-interest and compliance measures in time periods outside of the period-of-interest. (fTCD#6rw, 61.6%).

Data exclusion

- Follow clear objective criteria for removal of outlier trials when the LI is calculated by averaging over several trials or epochs. (fTCDS#11, 100% agreement).

Data processing

- An analysis pipeline for fTCD data should include down sampling, normalization, heart cycle integration, epoching, data screening, artifact rejection, and baseline correction. (fTCD#12rw, 100% agreement).

- Perform normalization and baseline correction on a trial-by-trial basis to avoid slow changes in blood flow velocity in one or both channels which can contaminate the results. (fTCD#13rw, 76% agreement).

Timings

- Regarding the period of interest (POI), develop standard methods and agree on an appropriate POI. If using the same task as used in a prior study, the same POI should be used, unless clear justification is given for using a different one. (fTCD#20, 100% agreement).
- During the development of new tasks, pilot-test the optimal duration of the task. (fTCD#18rw, 84.7% agreement).

Reporting

- When sharing raw fTCD data, include a file with analysis parameters, such as trial onset and end time, trigger channels, number of trials, etc. (fTCDS#32, 100% agreement).
- When developing new language tasks, present their strength of lateralization together with LIs derived from a fluency task standardized across the field. (fTCD#34rw, 100% agreement).
- When describing the results of an fTCDS experiment, report:
    - Both the direction and degree of the LI (fTCD#27, 100% agreement).
    - The confidence interval of the LI. (fTCD#29, 92.3% agreement).
    - An index of variability across trials and a visualization showing individual Lis for each task. (fTCD#30, 90% agreement).
    - The grand average of the fTCD cerebral flow velocity change relative to the baseline (fTCD#28, 84.6% agreement).
- It is recommended that the methods section of an fTCD experiment includes:
    - Criteria for excluding or terminating a fTCD recording. It is advised these criteria are determined a priori (fTCD#10, 100% agreement)
    - Whether the LI was based on the mean difference between left and right channels or the peak difference. The choice for either method should be justified. (fTCD#14, 100% agreement).
    - The period of interest. (fTCD#18rw, 84.7% agreement).
    - The number of trials per epoch presented per condition and the cut-off for excluding participants for too few trials. (fTCD#26, 100% agreement).
    - A set of participant characteristics (including biological sex, age, handedness, history of neurological disorders, degree of bilingualism/multilingualism and possibly medications) as well as how participants were recruited. (fTCD#23rw, 92.3% agreement).
    - The maximum gain and power settings using during the Doppler recording (fTCD#22, 76.9%).
- It is recommended to routinely supply anonymized data files with the manuscript. (fTCD#31, 77% agreement).
- It is advised to report a participant's cognitive abilities relevant to the task. (fTCD#24, 53.9%).

Critical review by Nicolas Badcock and Dorothy Bishop

Study set-up. While most experts agreed that the LI is likely to be unreliable if based on fewer than 16 trials, some commenters wondered on which criteria this was based and suggested that this minimum number may depend on the task at hand. One way to determine the optimal set-up is to pilot a task with, say, 20 trials, and then analyse the data with different subsets of trials, to see how this affects laterality estimates. Because the formula for a standard error of an estimate is divided by the square root of N, precision of measurement tends to increase substantially with more trials when N is small, and then more gradually as N trials increases. For instance, if the standard deviation of the measure is 1, then the standard error will be .50, .35, .29, .25, and .22 for trial N of 4, 8, 12, 16 and 20 respectively.

There was clear support for analysing participants' behaviour during the period of interest, but one commenter warned that overt/active tasks may induce noise and reduce the reliability of the measurement. The importance of assessing motor responses during fTCD tasks was acknowledged, but comments stated doing so comes with an additional burden and may not always be feasible. A suggested alternative was to balance left and right responses to avoid asymmetrical motor activation from contaminating the measurement.

Data handling and task design. A common sentiment between comments across statements was that more research is needed to examine how decisions regarding task design and data pre-processing may affect laterality and its reliability, for example the use of overt tasks, the inclusion of compliance measures, or the use of trial-by-trial normalization. Another recurring theme was that commenters in principle agreed with suggestions made by statements but admitted that implementing them would be challenging. For instance, deciding on design parameters of new paradigms by thoroughly piloting them may not be possible due to limited resources, so that reaching a consensus about the preferred timeframe around the duration of baseline and normalization periods may be unfeasible.

Reporting. This is a relatively young area of research, and there is still a lot to learn about how methodological decisions affect findings. It would be unwise to be too prescriptive about the design and analysis of studies in this area, but it is useful to list the factors that need to be considered and reported when designing and analysing laterality experiments using fTCD. These include:

- Number of trials administered and included (fTCD#26)
- Whether or not presentation of different trial types is blocked (fTCD#26)
- Whether behaviour is recorded during a period of interest (fTCD#5)
- Compliance with task demands (fTCD#6rw)
- Motor activity during the task (fTCD#7rw)
- Criteria for excluding participants and trials (fTCD#10, fTCD#11)
- Analysis pipeline (fTCD#12)
- How the laterality index was computed (fTCD#14, fTCD#27)
- Duration of epoch, and salient timepoints (fTCD#18rw, fTCD#19)
- Technical settings (e.g., gain, power settings) (fTCD#22)
- Characteristics of participants (fTCD#23rw, fTCD#24)

As with other sections, given the uncertainty about best practice, to make progress we need to encourage open sharing of raw data (fTCD#31). This would mean that researchers can, if they wish, reanalyse data to explore the impact of different analytic decisions on findings.

For example, a recent paper by Thompson et al (2022) compared traditional ways of computing the laterality index from fTCD data with more sophisticated methods based on fMRI, including use of Generalized Linear Models and Generalized Additive Models. The same datasets were analysed with each approach, showing that although the mean estimates of LI were closely similar across methods, the precision of the LI estimate was much greater when Generalized Additive Models were used. The data from this study are openly available, so that other researchers could explore issues such as the optimal number of trials to give reliable estimates.

In addition, preregistration of study protocol (introduction, methods, and analysis plan) is desirable to make it possible to distinguish confirmatory from exploratory findings. Finally, the reviewers recommend EQUATOR reporting standards for quantitative research provide checklists for a range of study designs; adherence to these standards would improve the reproducibility of research in this field. https://www.equator-network.org/reporting-guidelines/journal-article-reporting-standards-for-quantitative-research-in-psychology-the-apa-publications-and-communications-board-task-force-report/

Outstanding issues

Although the majority of experts agreed that the LI is likely to be unreliable if fewer than 16 trials are presented (fTCDS #1rw), several wonder why this cut-off was proposed to be 16 specifically. It may be worthwhile doing pilot studies to compare the impact of using different numbers of trials for a given task.

Several statements that had a high degree of agreement call for standardization across studies, including reporting a standard set of participant characteristics (fTCD #23), trial timings for widely-used tasks (fTCD #18), and a preferred time-frame around the duration of the baseline and normalization periods for different fTCD tasks (fTCD #19). It will take further research to decide what these standardizations should be.

General discussion

General observations

Consensus comes from the Latin word *consentire,* literally meaning 'feeling together', and refers to a general agreement in opinion. Definitions as to the required level of agreement to achieve consensus vary from majority to unanimity, which brings the range of required votes from plus 50 to 100% of experts. A review of a random sample of 100 English language Delphi studies revealed that definitions of consensus vary widely and are poorly reported (Diamond et al., 2014). Here, we used conventional majority as the leading principle to retain statements for the next round and to list them as recommendations for good practise. While it can be debated if a conventional majority would qualify as the required level of consensus we are looking for, it provides an established principle in democratic decision-making based on votes.

When it comes to statements about more technical aspects of measurement, apparatus setup, or validity checks, consensus ratings are generally distinct, and we retained those to remind the researcher of what is considered good practise by the majority of an expert panel. In the general and preference/performance sections, opinions are more equivocal. While experts appear eager to obtain consensus on a general laterality index formula and its scale, on a standard handedness inventory, on whether to include bimanual trials, and on how to conceptualize mixed/bilateral handedness, no single option received a majority of the votes. MayI is naive to expect consensus from a simple list of multiple-choice items that does not provide much argument or discussion, let alone empirical data to back it up. Perhaps the right answer was not in the list. Rather than being disappointed by this (lack of) result, we should prioritize the finding that a fair majority of experts are willing to strive for a consensus on these matters but that a more focused and empirically supported effort will be necessary to achieve it. In general, it could be argued that, in the face of divisions in expert opinion, we should let researchers continue with their favourite approach, so long as it is clearly described. Nevertheless, there are cases where measurement choices, such as the method of defining the LI in behavioural measures, have little practical impact, and adoption of a consistent approach would be hugely beneficial to facilitating communication in the field.

Another general observation is the disunity of the expert panel with regard to the interpretation of 'guidelines'. While some appear to embrace a clear set of recommendations that researchers could apply to make their findings more comparable and transparent, others interpret them as a set of rules and shiver at the idea that a prescribed set of guidelines would ultimately become a reviewer's checklist of requirements for publication. In the latter case, guidelines may be detrimental to academic freedom and original scientific practise, and this, of course, is not what we had in mind with this survey. Although we explicitly mentioned this in our communication, the concern that guidelines may evolve into prescriptions that may not fit every research question is mentioned repeatedly, and this sentiment should be taken seriously by any attempt to promote more uniformity in laterality research. We therefore refrained from using the term 'guidelines' in this paper and used 'recommendations' instead.

Study limitations

More than 100 researchers (with near parity between female and male researchers), most of them in senior positions and with solid academic references, have contributed to the LICI project. Some continents are underrepresented (e.g., South America, Asia) or not represented at all (e.g., Africa). It should also be noted that while all experts completed the General section, experts were next invited to complete the sections of their expertise, which implies that some sections like the electrophysiology and fTCD sections, were completed by 20 or fewer experts. No or not enough statements regarding fNIRS and functional connectivity (e.g., resting state, connectome) were submitted to warrant a separate section.

While the LICI aimed to focus on the use of laterality indices this was broadly interpreted by the expert panel. Strictly speaking, conceptual issues regarding ambidexterity, egocentric space, and bilateral language representation, to name a few, have little to do with laterality indices, yet these types of questions turned up, and not only in the general section. We decided to keep these statements in the survey, as they reflect concerns researchers of laterality run into. The drawback of this approach is that experts were required to address many diverse and complex issues in a single survey which made the task (much) more demanding than originally anticipated.

A final limitation is that our method did not allow for an exchange of views on the more complex topics. As the number of Rounds was limited and the list of statements long, 'discussion' was limited to a vote and a comment box and no real interaction took place. While this strategy was sufficient for most technical statements, our approach was too superficial to tackle more controversial or general issues.

The way forward

A majority agreement was reached for a number of statements, which allowed compilation of an (nonexhaustive) set of recommendations. Other matters remain outstanding and could serve as a starting point for future endeavours. Throughout the LICI, the laterality community has proven itself keen to work together on common issues and our hope is that such collaborative efforts will continue to be launched in years to come.

Despite the limitations mentioned above, we argue that the Delphi-approach continues to offer a contemporary and structured method to explore consensus among experts. In hindsight, some recommendations can be made:

- Select one well-defined topic (and stick to it)
- Carefully prepare a structured and peer-reviewed survey
- Invite a large number of experts to collaborate on this topic
- Collect votes and comments in a first round and identify bottlenecks
- Finetune statements in consecutive rounds and feedback (empirical) arguments for divergent opinions
- Report the outcome when a status quo in the voting results is obtained

Including follow-up (Delphi) surveys, future initiatives could take the form of formal consensus meetings about specific issues, meetings to tackle problems collaboratively in a hands-on, data-d–iven fashion - akin to a "laterality-specific brainhack" (Gau et al., 2021) as suggested by one of the commenters, and data-sharing as well as running multi-site studies to fill in gaps that stand in the way of formulating evidence-based best practice recommendations.

List of contributing experts (alphabetical order)

All experts mentioned here consented to be named as contributors to the LICI project.

John Allen, Ruth Atchley, Tatjana Aue, Monica Baciu, Nicholas Badcock, Marie Banich, Alan Beaton, Christopher Benjamin, Dorothy Bishop, Pamela Bryden, Marc Brysbaert, Karen Caeyenberghs, Julie Campbell, David Carey, Nicolas Cherbuin, Steve Christman, Helene Cochet, Daniela Corbetta, Patricia Cowell, Fabrice Crivello, Erik Domellöf, Jessica Dubois, Lisa Eyler, Jacqueline Fagard, Jason Flindall, Gillian Forrester, Nathan Fox, Clyde Francks, Patrick Friedrich, Jurgen Germann, Robin Gerrits, Reint Geuze, Anna Grabowska, Margriet Groen, Eva Gutierrez-Sigut, Scott Hardie, Eddie Harmon-Jones, Lauren Julius Harris, Markus Hausmann, Johannes Hewig, Marco Hirnstein, Jessica Hodgson, Kenneth Hugdahl, Vasiliki Iliadou, Lutz Jäncke, Marc Joliot, Shogo Kajimura, Jurgen Kayser, Xiang-Zhen Kong, Dimitrios Kourtis, Gregory Kroliczak, Kristina Kuper, Rotem Leshem, Hesheng Liu, Eileen Luders, Jessica Lust, Keith Lyle, Mairead MacSweeney, Jean-François Mangin, Emily Marcinowski, Alexandre Marcori, Emmanuel Mellet, Sanja Milenkovic, Christine Mohr, Martin Musalek, Eliza Nelson, Mike Nicholls, Sebastian Ocklenburg, Victor Okazaki, Matia Okubo, Marietta Papadatou-Pastou, Ilona Papousek, Silvia Paracchini, Heather Payne, Laurent Petit, Clare Porac, Giulia Prete, Jacques Prieur, Neil Roberts, Fabrice Sarlegna, Astrid Schepman, Judith Schmitz, Mohamed Seghier, Deborah Serrien, Gabriele Soffritti, Karsten Specht, Jerzy Szaflarski, Michel Thiebaut de Schotten, Mattie Tops, Ulrich Tran, Natalie Uomini, Jyotsna Vaid, Helena Verhelst, Guy Vingerhoets, Daniel Voyer, Kate Watkins, Rene Westerhausen, Marko Wilke, Zoe Woodhead, Lynn Wright, Laure Zago.

References

Adcock, J. E., Wise, R. G., Oxbury, J. M., Oxbury, S. M., & Matthews, P. M. (2003). Quantitative fMRI assessment of the differences in lateralization of language-related brain activation in patients with temporal lobe epilepsy. *Neuroimage*, *18*(2), 423-438. https://doi.org/10.1016/s1053-8119(02)00013-7

Aghamollaei, M., Jafari, Z., Tahaei, A., Toufan, R., Keyhani, M., Rahimzade, S., & Esmaeili, M. (2013). Dichotic assessment of verbal memory function: development and validation of the Persian version of Dichotic Verbal Memory Test. *Journal of the American Academy of Audiology*, *24*(8), 684-688.

Allen, J. J. B., Coan, J. A., & Nazarian, M. (2004). Issues and assumptions on the road from raw signals to metrics of frontal EEG asymmetry in emotion. *Biological Psychology*, *67*(1-2), 183-218. https://doi.org/10.1016/j.biopsycho.2004.03.007

Asbjørnsen, A. E., & Bryden, M. (1996). Biased attention and the fused dichotic words test. *Neuropsychologia*, *34*(5), 407-411.

Baciu, M., Juphard, A., Cousin, E., & Le Bas, J. F. (2005). Evaluating fMRI methods for assessing hemispheric language dominance in healthy subjects. *European Journal of Radiology*, *55*(2), 209-218. https://doi.org/10.1016/j.ejrad.2004.11.004

Bailey, L. M., McMillan, L. E., & Newman, A. J. (2020). A sinister subject: Quantifying handedness-based recruitment biases in current neuroimaging research. *Eur J Neurosci*, *51*(7), 1642-1656. https://doi.org/10.1111/ejn.14542

Beaumont, J. G. (1982). Studies with verbal stimuli. In J. D. Beaumont (Ed.), *Divided visual field studies of cerebral organisation* (pp. 57-86). Academic Press.

Benjamin, C. F., Walshaw, P. D., Hale, K., Gaillard, W. D., Baxter, L. C., Berl, M. M., Polczynska, M., Noble, S., Alkawadri, R., Hirsch, L. J., Constable, R. T., & Bookheimer, S. Y. (2017). Presurgical language fMRI: Mapping of six critical regions. *Hum Brain Mapp*, *38*(8), 4239-4255. https://doi.org/10.1002/hbm.23661

Berlin, C. I., Lowe-Bell, S. S., Cullen Jr, J. K., Thompson, C. L., & Stafford, M. R. (1972). Is speech "special"? Perhaps the temporal lobectomy patient can tell us. *The Journal of the Acoustical Society of America*, *52*(2B), 702-705.

Bishop, D. V. M. (1989). Does Hand Proficiency Determine Hand Preference. *British Journal of Psychology*, *80*, 191-199. https://doi.org/DOI 10.1111/j.2044-8295.1989.tb02313.x

Bishop, D. V. M. (1990a). *Handedness and Developmental Disorder*. MacKeith Press.

Bishop, D. V. M. (1990b). How to Increase Your Chances of Obtaining a Significant Association between Handedness and Disorder. *Journal of Clinical and Experimental Neuropsychology*, *12*(5), 812-816. https://doi.org/Doi 10.1080/01688639008401022

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & Consortium, C. (2016). CATALISE: A Multinational and Multidisciplinary Delphi Consensus Study. Identifying Language Impairments in Children (vol 11, e0158753, 2016). *Plos One*, *11*(12). https://doi.org/ARTN e0168066 10.1371/journal.pone.0168066

Boklage, C. E. (1980). The Sinistral Blastocyst: An Embryologic Perspective on the Development of Brain-Function Asymmetries. In J. Herron (Ed.), *Neuropsychology of Left-Handedness* (pp. 115-137). Academic Press.

Boulkedid, R., Abdoul, H., Loustau, M., Sibony, O., & Alberti, C. (2011). Using and Reporting the Delphi Method for Selecting Healthcare Quality Indicators: A Systematic Review. *Plos One*, *6*(6). https://doi.org/ARTN e20476, 10.1371/journal.pone.0020476

Bourne, V. J. (2006). The divided visual field paradigm: Methodological considerations. *Laterality*, *11*(4), 373-393. https://doi.org/10.1080/13576500600633982

Bradshaw, A. R., Bishop, D. V. M., & Woodhead, Z. V. J. (2017). Methodological considerations in assessment of language lateralisation with fMRI: a systematic review. *Peerj*, *5*. https://doi.org/ARTN e3557 0.7717/peerj.3557

Branco, D. M., Suarez, R. O., Whalen, S., O'Shea, J. P., Nelson, A. P., da Costa, J. C., & Golby, A. J. (2006). Functional MRI of memory in the hippocampus: Laterality indices may be more meaningful if calculated from whole voxel distributions. *Neuroimage*, *32*(2), 592-602. https://doi.org/10.1016/j.neuroimage.2006.04.201

Brown, S. G., Roy, E. A., Rohr, L. E., Snider, B. R., & Bryden, P. J. (2004). Preference and performance measures of handedness. *Brain and Cognition*, *55*(2), 283-285. https://doi.org/10.1016/j.bandc.2004.02.010

Bryden, M. (1962). Order of report in dichotic listening. *Canadian Journal of Psychology*, *16*, 291-299.

Bryden, M. P., & Sprott, D. A. (1981). Statistical Determination of Degree of Laterality. *Neuropsychologia*, *19*(4), 571-581. https://doi.org/Doi 10.1016/0028-3932(81)90023-3

Bryden, P. J., Pryde, K. M., & Roy, E. A. (2000). A developmental analysis of the relationship between hand preference and performance: II. A performance-based method of measuring hand preference in children. *Brain and Cognition*, *43*(1-3), 60-64. <Go to ISI>://WOS:000087157600042

Buenaventura Castillo, C., Lynch, A. G., & Paracchini, S. (2020). Different laterality indexes are poorly correlated with one another but consistently show the tendency of males and females to be more left- and right-lateralized, respectively. *Royal Society Open Science*, *7*(4), 191700. https://doi.org/doi:10.1098/rsos.191700

Burle, B., Spieser, L., Roger, C., Casini, L., Hasbroucq, T., & Vidal, F. (2015). Spatial and temporal resolutions of EEG: Is it really black and white? A scalp current density view. *International Journal of Psychophysiology*, *97*(3), 210-220. https://doi.org/10.1016/j.ijpsycho.2015.05.004

Busch, D., Hagemann, N., & Bender, N. (2010). The dimensionality of the Edinburgh Handedness Inventory: An analysis with models of the item response theory. *Laterality*, *15*(6), 610-628. https://doi.org/10.1080/13576500903081806

Campbell, J. M., Marcinowski, E. C., Latta, J., & Michel, G. F. (2015). Different assessment tasks produce different estimates of handedness stability during the eight to 14 month age period. *Infant Behavior & Development*, *39*, 67-80. https://doi.org/10.1016/j.infbeh.2015.02.003

Carson, R. G., Goodman, D., Chua, R., & Elliott, D. (1993). Asymmetries in the regulation of visually guided aiming. *J Mot Behav*, *25*(1), 21-32. https://doi.org/10.1080/00222895.1993.9941636

Cherry, B. J., Hellige, J. B., & McDowd, J. M. (1995). Age differences and similarities in patterns of cerebral hemispheric asymmetry. *Psychol Aging*, *10*(2), 191-203. https://doi.org/10.1037//0882-7974.10.2.191

Chlebus, P., Mikl, M., Brazdil, M., Pazourkova, M., Krupa, P., & Rektor, I. (2007). fMRI evaluation of hemispheric language dominance using various methods of laterality index calculation. *Experimental Brain Research*, *179*(3), 365-374. https://doi.org/10.1007/s00221-006-0794-y

Cohen, M. E., & Ross, L. E. (1977). Saccade Latency in Children and Adults - Effects of Warning Interval and Target Eccentricity. *Journal of Experimental Child Psychology*, *23*(3), 539-549. https://doi.org/Doi 10.1016/0022-0965(77)90044-3

Cohen, M. X. (2015). Comparison of different spatial transformations applied to EEG data: A case study of error processing. *International Journal of Psychophysiology*, *97*(3), 245-257. https://doi.org/10.1016/j.ijpsycho.2014.09.013

Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. *Psychological Review*, *83*(2), 114.

De Schryver, M., Hughes, S., Rosseel, Y., & De Houwer, J. (2016). Unreliable Yet Still Replicable: A Comment on LeBel and Paunonen (2011) [Methods]. *Frontiers in Psychology*, *6*(2039). https://doi.org/10.3389/fpsyg.2015.02039

Diamond, I. R., Grant, R. C., Feldman, B. M., Pencharz, P. B., Ling, S. C., Moore, A. M., & Wales, P. W. (2014). Defining consensus: A systematic review recommends methodologic criteria for reporting of Delphi studies. *Journal of Clinical Epidemiology*, *67*(4), 401-409. https://doi.org/10.1016/j.jclinepi.2013.12.002

Dragovic, M. (2004). Towards an improved measure of the Edinhurgh Handedness Inventory: A one-factor congeneric measurement model using confirmatory factor analysis. *Laterality*, *9*(4), 411-419. https://doi.org/10.1080/13576500342000248

Edlin, J. M., Leppanen, M. L., Fain, R. J., Hacklander, R. P., Hanaver-Torrez, S. D., & Lyle, K. B. (2015). On the use (and misuse?) of the Edinburgh Handedness Inventory. *Brain Cogn*, *94*, 44-51. https://doi.org/10.1016/j.bandc.2015.01.003

Fagard, J., Margules, S., Lopez, C., Granjon, L., & Huet, V. (2017). How should we test infant handedness? *Laterality*, *22*(3), 294-312. https://doi.org/10.1080/1357650x.2016.1192186

Fazio, R., Coenen, C., & Denney, R. L. (2012). The original instructions for the Edinburgh Handedness Inventory are misunderstood by a majority of participants. *Laterality*, *17*(1), 70-77. https://doi.org/10.1080/1357650x.2010.532801

Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., & Elger, C. E. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology*, *60*(6), 969-975. https://doi.org/10.1212/01.Wnl.0000049934.34209.2e

Fesl, G., Bruhns, P., Rau, S., Wiesmann, M., Ilmberger, J., Kegel, G., & Brueckmann, H. (2010). Sensitivity and reliability of language laterality assessment with a free reversed association task-a fMRI study. *European Radiology*, *20*(3), 683-695. https://doi.org/10.1007/s00330-009-1602-4

Fiander, M., & Burns, T. (1998). Essential components of schizophrenia care: a Delphi approach. *Acta Psychiatr Scand*, *98*(5), 400-405. https://doi.org/10.1111/j.1600-0447.1998.tb10105.x

Fischer, B., & Ramsperger, E. (1984). Human Express Saccades - Extremely Short Reaction-Times of Goal Directed Eye-Movements. *Experimental Brain Research*, *57*(1), 191-195. <Go to ISI>://WOS:A1984TS47800020

Fischer, B., & Weber, H. (1993). Express Saccades and Visual-Attention. *Behavioral and Brain Sciences*, *16*(3), 553-567. https://doi.org/Doi 10.1017/S0140525x00031575

Flowers, K. (1975). Handedness and Controlled Movement. *British Journal of Psychology*, *66*(Feb), 39-52. https://doi.org/DOI 10.1111/j.2044-8295.1975.tb01438.x

Freides, D. (1977). Do dichotic listening procedures measure lateralization of information processing or retrieval strategy? *Perception & Psychophysics*, *21*(3), 259-263.

Friedrich, P., Ocklenburg, S., Mochalski, L., Schluter, C., Gunturkun, O., & Genc, E. (2017). Long-term reliability of the visual EEG Poffenberger paradigm. *Behavioural Brain Research*, *330*, 85-91. https://doi.org/10.1016/j.bbr.2017.05.019

Gau, R., Noble, S., Heuer, K., Bottenhorn, K. L., Bilgin, I. P., Yang, Y. F., Huntenburg, J. M., Bayer, J. M. M., Bethlehem, R. A. I., Rhoads, S. A., Vogelbacher, C., Borghesani, V., Levitis, E., Wang, H. T., Van den Bossche, S., Kobeleva, X., Legarreta, J. H., Guay, S., Atay, S. M., Varoquaux, G. P., Huijser, D. C., Sandstrom, M. S., Herholz, P., Nastase, S. A., Badhwar, A., Dumas, G., Schwab, S., Moia, S., Dayan, M., Bassil, Y., Brooks, P. P., Mancini, M., Shine, J. M., O'Connor, D., Xie, X. H., Poggiali, D., Friedrich, P., Heinsfeld, A. S., Riedl, L., Toro, R., Caballero-Gaudes, C., Eklund, A., Garner, K. G., Nolan, C. R., Demeter, D. V., Barrios, F. A., Merchant, J. S., McDevitt, E. A., Oostenveld, R., Craddock, R. C., Rokem, A., Doyle, A., Ghosh, S. S., Nikolaidis, A., Stanley, O. W., Urunuela, E., & Community, B. (2021). Brainhack: Developing a culture of open, inclusive, community-driven neuroscience. *Neuron*, *109*(11), 1769-1775. https://doi.org/10.1016/j.neuron.2021.04.001

Geffen, G., Bradshaw, J. L., & Nettleton, N. C. (1972). Hemispheric Asymmetry - Verbal and Spatial Encoding of Visual Stimuli. *Journal of Experimental Psychology*, *95*(1), 25-+. https://doi.org/DOI 10.1037/h0033265

Goble, D. J., & Brown, S. H. (2007). Task-dependent asymmetries in the utilization of proprioceptive feedback for goal-directed movement. *Experimental Brain Research*, *180*(4), 693-704. https://doi.org/10.1007/s00221-007-0890-7

Goodman, C. M. (1987). The Delphi technique: a critique. *Journal of Advanced Nursing*, *12*(6), 729-734. https://doi.org/10.1111/j.1365-2648.1987.tb01376.x

Grissom, R. J., & Kim, J. J. (2012). *Effect sizes for research: Univariate and multivariate applications*. Routledge.

Guiard, Y. (1987). Asymmetric Division of Labor in Human Skilled Bimanual Action - the Kinematic Chain as a Model. *Journal of Motor Behavior*, *19*(4), 486-517. <Go to ISI>://WOS:A1987N201900005

Hardie, S. M., & Wright, L. (2014). Differences between left- and right-handers in approach/avoidance motivation: influence of consistency of handedness measures. *Front Psychol*, *5*, 134. https://doi.org/10.3389/fpsyg.2014.00134

Hasson, F., Keeney, S., & McKenna, H. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing*, *32*(4), 1008-1015. https://doi.org/DOI 10.1046/j.1365-2648.2000.01567.x

Hausmann, M., Becker, C., Gather, U., & Gunturkun, O. (2002). Functional cerebral asymmetries during the menstrual cycle: a cross-sectional and longitudinal analysis. *Neuropsychologia*, *40*(7), 808-816. https://doi.org/10.1016/S0028-3932(01)00179-8

Hausmann, M., & Gunturkun, O. (1999). Sex differences in functional cerebral asymmetries in a repeated measures design. *Brain and Cognition*, *41*(3), 263-275. https://doi.org/DOI 10.1006/brcg.1999.1126

Hausmann, M., Hodgetts, S., & Eerola, T. (2016). Music-induced changes in functional cerebral asymmetries. *Brain and Cognition*, *104*, 58-71. https://doi.org/10.1016/j.bandc.2016.03.001

Hausmann, M., Kirk, I. J., & Corballis, M. C. (2004). Influence of task complexity on manual asymmetries. *Cortex*, *40*(1), 103-110. https://doi.org/Doi 10.1016/S0010-9452(08)70923-7

Hiscock, M., Cole, L. C., Benthall, J. G., Carlson, V. L., & Ricketts, J. M. (2000). Toward solving the inferential problem in laterality research: Effects of increased reliability on the validity of the dichotic listening right-ear advantage. *Journal of the International Neuropsychological Society*, *6*(5), 539-547.

Hugdahl, K., Westerhausen, R., Alho, K., Medvedev, S., & Hämäläinen, H. (2008). The effect of stimulus intensity on the right ear advantage in dichotic listening. *Neuroscience letters*, *431*(1), 90-94.

Hugdahl, K., Westerhausen, R., Alho, K., Medvedev, S., Laine, M., & Hämäläinen, H. (2009). Attention and cognitive control: unfolding the dichotic listening story. *Scandinavian journal of psychology*, *50*(1), 11-22.

Hunter, Z. R., & Brysbaert, M. (2008). Visual half-field experiments are a good measure of cerebral language dominance if used properly: Evidence from fMR1. *Neuropsychologia*, *46*(1), 316-325. <Go to ISI>://000253362300031 (Not in File)

Iliadou, V., Bamiou, D. E., Kaprinis, S., Kandylis, D., & Kaprinis, G. (2009). Auditory Processing Disorders in children suspected of Learning Disabilities-A need for screening? *International Journal of Pediatric Otorhinolaryngology*, *73*(7), 1029-1034. https://doi.org/10.1016/j.ijporl.2009.04.004

Iliadou, V., Ptok, M., Grech, H., Pedersen, E. R., Brechmann, A., Deggouj, N., Kiese-Himmel, C., Sliwinska-Kowalska, M., Nickisch, A., Demanez, L., Veuillet, E., Hung, T. V., Sirimanna, T., Callimachou, M., Santarelli, R., Kuske, S., Barajas, J., Hedjever, M., Konukseven, O., Veraguth, D., Mattsson, T. S., Martins, J. H., & Bamiou, D. E. (2017). A European Perspective on Auditory Processing Disorder-Current Knowledge and Future Research Focus. *Frontiers in Neurology*, *8*. https://doi.org/ARTN 622 10.3389/fneur.2017.00622

Jansen, A., Menke, R., Sommer, J., Forster, A. F., Bruchmann, S., Hempleman, J., Weber, B., & Knecht, S. (2006). The assessment of hemispheric lateralization in functional MRI - Robustness and reproducibility. *Neuroimage*, *33*(1), 204-217. <Go to ISI>://000241209800022

Jayasinghe, S. A. L., Sarlegna, F. R., Scheidt, R. A., & Sainburg, R. L. (2021). Somatosensory deafferentation reveals lateralized roles of proprioception in feedback and adaptive feedforward control of movement and posture. *Current Opinion in Physiology*, *19*, 141-147. https://doi.org/10.1016/j.cophys.2020.10.005

Johnstone, L. T., Karlsson, E. M., & Carey, D. P. (2020). The validity and reliability of quantifying hemispheric specialisation using fMRI: Evidence from left and right handers on three different cerebral asymmetries. *Neuropsychologia*, *138*, 107331. https://doi.org/10.1016/j.neuropsychologia.2020.107331

Joliot, M., Jobard, G., Naveau, M., Delcroix, N., Petit, L., Zago, L., Crivello, F., Mellet, E., Mazoyer, B., & Tzourio-Mazoyer, N. (2015). AICHA: An atlas of intrinsic connectivity of homotopic areas. *Journal of Neuroscience Methods*, *254*, 46-59. https://doi.org/10.1016/j.jneumeth.2015.07.013

Kayser, J., & Tenke, C. E. (2003). Optimizing PCA methodology for ERP component identification and measurement: theoretical rationale and empirical evaluation. *Clinical Neurophysiology*, *114*(12), 2307-2325. https://doi.org/10.1016/S1388-2457(03)00241-4

Kayser, J., & Tenke, C. E. (2005). Trusting in or breaking with convention: Towards a renaissance of principal components analysis in electrophysiology. *Clinical Neurophysiology*, *116*(8), 1747-1753. https://doi.org/10.1016/j.clinph.2005.03.020

Kayser, J., & Tenke, C. E. (2015a). Hemifield-dependent N1 and event-related theta/delta oscillations: An unbiased comparison of surface Laplacian and common EEG reference choices. *International Journal of Psychophysiology*, *97*(3), 258-270. https://doi.org/10.1016/j.ijpsycho.2014.12.011

Kayser, J., & Tenke, C. E. (2015b). Issues and considerations for using the scalp surface Laplacian in EEG/ERP research: A tutorial review. *International Journal of Psychophysiology*, *97*(3), 189-209. https://doi.org/10.1016/j.ijpsycho.2015.04.012

Kimura, D. (1961). Some effects of temporal-lobe damage on auditory perception. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *15*(3), 156-165.

Kompus, K., Falkenberg, L. E., Bless, J. J., Johnsen, E., Kroken, R. A., Kråkvik, B., Larøi, F., Løberg, E.-M., Vedul-Kjelsås, E., Westerhausen, R., & Hugdahl, K. (2013). The role of the primary auditory cortex in the neural mechanism of auditory verbal hallucinations. *Frontiers in Human Neuroscience*, *7*, 144.

Labache, L., Mazoyer, B., Joliot, M., Crivello, F., Hesling, I., & Tzourio-Mazoyer, N. (2020). Typical and atypical language brain organization based on intrinsic connectivity and multitask functional asymmetries. *Elife*, *9*. https://doi.org/ARTN e58722 10.7554/eLife.58722

Leppanen, M. L., Lyle, K. B., Edlin, F. M., & Schafke, V. D. (2019). Is self-report a valid measure of unimanual object-based task performance? *Laterality*, *24*(5), 538-558. https://doi.org/10.1080/1357650x.2018.1550493

Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs(?). *Behavior Research Methods*, *51*(1), 40-60. https://doi.org/10.3758/s13428-018-1076-x

Linstone, H. A., & Turoff, M. (1975). *The Delphi method : techniques and applications*. Addison-Wesley Pub. Co., Advanced Book Program.

Lyle, K. B., Roediger, H. L., & McCabe, D. P. (2008). Handedness is related to memory via hemispheric interaction: Evidence from paired associate recall and source memory tasks. *Neuropsychology*, *22*(4), 523-530. https://doi.org/10.1037/0894-4105.22.4.523

Marim, E. d. A. L., R.; Okazaki, V.H.A. (2011). Global lateral preference inventory. *Brazilian Journal of Motor Behavior*, *6*(3), 14–23. https://doi.org/https://doi.org/10.20338/bjmb.v6i3.178

Mathew, J., Sarlegna, F. R., Bernier, P.-M., & Danion, F. R. (2019). Handedness Matters for Motor Control But Not for Prediction. *eneuro*, *6*(3), ENEURO.0136-0119.2019. https://doi.org/10.1523/eneuro.0136-19.2019

Matsuo, K., Kono, K., Shimoda, K., Kaji, Y., & Akiyama, K. (2021). Reproducibility of the lateralization index in functional magnetic resonance imaging across language tasks. *Journal of Neurolinguistics*, *57*. https://doi.org/ARTN 100943 10.1016/j.jneuroling.2020.1009 3

Mckenna, H. P. (1994). The Delphi Technique - a Worthwhile Research Approach for Nursing. *Journal of Advanced Nursing*, *19*(6), 1221-1225. https://doi.org/DOI 10.1111/j.1365-2648.1994.tb01207.x

McManus, I. C., Van Horn, J. D., & Bryden, P. J. (2016). The Tapley and Bryden test of performance differences between the hands: The original data, newer data, and the relation to pegboard and other tasks. *Laterality*, *21*(4-6), 371-396. https://doi.org/10.1080/1357650x.2016.1141916

Milenkovic, S., & Dragovic, M. (2013). Modification of the Edinburgh Handedness Inventory: A replication study. *Laterality*, *18*(3), 340-348. https://doi.org/10.1080/1357650x.2012.683196

Mohr, C., Michel, C. M., Lantz, G., Ortigue, S., Viaud-Delmon, L., & Landis, T. (2005). Brain state-dependent functional hemispheric specialization in men but not in women. *Cerebral Cortex*, *15*(9), 1451-1458. https://doi.org/10.1093/cercor/bhi025

Musalek, M. (2014). *Development of TestBatteries for Diagnostics of Motor Laterality Manifestation – Link between Cerebellar Dominance and Hand Performance*. Karolinum Press.

Musalek, M. (2015). Skilled performance tests and their use in diagnosing handedness and footedness at children of lower school age 8-10. *Frontiers in Psychology*, *5*. https://doi.org/ARTN 1513 10.3389/fpsyg.2014.01513

Musalek, M., Scharoun, S. M., & Bryden, P. J. (2015). The Link Between Cerebellar Dominance and Skilled Hand Performance in 8-10-Year-Old Right-Handed Children. *Journal of Motor Behavior*, *47*(5), 386-396. https://doi.org/10.1080/00222895.2014.1003778

Musiek, F. E. (1983). Assessment of central auditory dysfunction: the dichotic digit test revisited. *Ear and hearing*, *4*(2), 79-83.

Nagata, S., Uchimura, K., Hirakawa, W., & Kuratsu, J. (2001). Method for quantitatively evaluating the lateralization of linguistic function using functional MR imaging. *American Journal of Neuroradiology*, *22*(5), 985-991. <Go to ISI>://WOS:000168681600030

Ocklenburg, S., Friedrich, P., Schmitz, J., Schluter, C., Genc, E., Gunturkun, O., Peterburs, J., & Grimshaw, G. (2019). Beyond frontal alpha: investigating hemispheric asymmetries over the EEG frequency spectrum as a function of sex and handedness. *Laterality*, *24*(5), 505-524. https://doi.org/10.1080/1357650x.2018.1543314

Oldfield, R. C. (1971). The assessment and analysis of handedness. *Neuropsychologia*, *9*, 97-113.

Otzenberger, H., Gounot, D., Marrer, C., Namer, I. J., & Metz-Lutz, M. N. (2005). Reliability of individual functional MRI brain mapping of language. *Neuropsychology*, *19*(4), 484-493. https://doi.org/10.1037/0894-4105.19.4.484

Packheiser, J., Schmitz, J., Berretz, G., Carey, D. P., Paracchini, S., Papadatou-Pastou, M., & Ocklenburg, S. (2020). Four meta-analyses across 164 studies on atypical footedness prevalence and its relation to handedness. *Scientific Reports*, *10*(1). https://doi.org/ARTN 14501 10.1038/s41598-020-71478-w

Parker, A. J., Woodhead, Z. V. J., Thompson, P. A., & Bishop, D. V. M. (2021). Assessing the reliability of an online behavioural laterality battery: A pre-registered study. *Laterality*, *26*(4), 359-397. https://doi.org/10.1080/1357650x.2020.1859526

Parsons, S., Kruijt, A. W., & Fox, E. (2019). Psychological Science Needs a Standard Practice of Reporting the Reliability of Cognitive-Behavioral Measurements. *Advances in Methods and Practices in Psychological Science*, *2*(4), 378-395. https://doi.org/10.1177/2515245919879695

Passow, S., Westerhausen, R., Wartenburger, I., Hugdahl, K., Heekeren, H. R., Lindenberger, U., & Li, S.-C. (2012). Human aging compromises attentional control of auditory perception. *Psychology and Aging*, *27*(1), 99.

Pedhazur, E. J., & Schmelkin, L. P. (2013). *Measurement, design, and analysis: An integrated approach*. Psychology Press.

Penner, I.-K., Schläfli, K., Opwis, K., & Hugdahl, K. (2009). The role of working memory in dichotic-listening studies of auditory laterality. *J Clin Exp Neuropsychol*, *31*(8), 959-966.

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., & Oostenveld, R. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, *6*. https://doi.org/10.1038/s41597-019-0104-8

Perrier, P. F., S. (2015). Motor Equivalence in Speech Production. In M. R. Redford (Ed.), *The Handbook of Speech Production* (pp. .225-247). Wiley-Blackwell.

Pill, J. (1971). Delphi Method - Substance, Context - Critique an an Annotated Bibliography. *Socio-Economic Planning Sciences*, *5*(1), 57-71. https://doi.org/Doi 10.1016/0038-0121(71)90041-3

Polit, D. F. (2014). Getting serious about test-retest reliability: a critique of retest research and some recommendations. *Quality of Life Research*, *23*(6), 1713-1720. https://doi.org/10.1007/s11136-014-0632-9

Porac, C. C., S. (1981). *Lateral Preferences and Human Behavior*. Springer. https://doi.org/ https://doi.org/10.1007/978-1-4613-8139-6

Powell, C. (2003). The Delphi technique: myths and realities. *Journal of Advanced Nursing*, *41*(4), 376-382. https://doi.org/10.1046/j.1365-2648.2003.02537.x

Prichard, E., Propper, R. E., & Christman, S. D. (2013). Degree of handedness, but not direction, is a systematic predictor of cognitive performance. *Frontiers in Psychology*, *4*. https://doi.org/ARTN 9 10.3389/fpsyg.2013.00009

Prichard, E. C., Christman, S. D., & Walters, J. (2020). The Pen Is Not Always Mightier: Different Ways of Measuring Handedness With the Edinburgh Handedness Inventory Yield Different Handedness Conclusions. *Perceptual and Motor Skills*, *127*(5), 789-802. https://doi.org/Artn 0031512520927562 10.1177/0031512520927562

Prieur, J., Barbu, S., & Blois-Heulin, C. (2017). Assessment and analysis of human laterality for manipulation and communication using the Rennes Laterality Questionnaire. *Royal Society Open Science*, *4*(8). https://doi.org/ARTN 170035 10.1098/rsos.170035

Provins, K. A. (1956). Handedness and Skill. *Quarterly Journal of Experimental Psychology*, *8*(2), 79-95. https://doi.org/Doi 10.1080/17470215608416806

Ramsey, N. F., Sommer, I. E., Rutten, G. J., & Kahn, R. S. (2001). Combined analysis of language tasks in fMRI improves assessment of hemispheric dominance for language functions in individual subjects. *Neuroimage*, *13*(4), 719-733. https://doi.org/10.1006/nimg.2000.0722

Rennie, D. (1981). Consensus statements. *N Engl J Med*, *304*(11), 665-666. https://doi.org/10.1056/NEJM198103123041110

Repp, B. H. (1976). Identification of dichotic fusions. *The Journal of the Acoustical Society of America*, *60*(2), 456-469.

Repp, B. H. (1977). Measuring Laterality Effects in Dichotic-Listening. *Journal of the Acoustical Society of America*, *62*(3), 720-737. https://doi.org/Doi 10.1121/1.381584

Rigal, R. A. (1992). Which Handedness - Preference or Performance. *Perceptual and Motor Skills*, *75*(3), 851-866. https://doi.org/Doi 10.2466/Pms.75.7.851-866

Rutten, G. J. M., Ramsey, N. F., van Rijen, P. C., & van Veelen, C. W. M. (2002). Reproducibility of fMRI-determined language lateralization in individual subjects. *Brain and Language*, *80*(3), 421-437. https://doi.org/10.1006/brln.2001.2600

Sackman, H. (1975). *DELPHI CRITIQUE*.

Sainburg, R. L., & Schaefer, S. Y. (2004). Interlimb differences in control of movement extent. *J Neurophysiol*, *92*(3), 1374-1383. https://doi.org/10.1152/jn.00181.2004

Satz, P. (1977). Laterality Tests - Inferential Problem. *Cortex*, *13*(2), 208-212. https://doi.org/Doi 10.1016/S0010-9452(77)80010-5

Schaefer, S. Y., Haaland, K. Y., & Sainburg, R. L. (2009). Hemispheric specialization and functional impact of ipsilesional deficits in movement coordination and accuracy. *Neuropsychologia*, *47*(13), 2953-2966. https://doi.org/10.1016/j.neuropsychologia.2009.06.025

Schaffer, J. E., Maenza, C., Good, D. C., Przybyla, A., & Sainburg, R. L. (2020). Left hemisphere damage produces deficits in predictive control of bilateral coordination. *Experimental Brain Research*, *238*(12), 2733-2744. https://doi.org/10.1007/s00221-020-05928-2

Schneider, D., Beste, C., & Wascher, E. (2012). On the time course of bottom-up and top-down processes in selective visual attention: An EEG study. *Psychophysiology*, *49*(11), 1660-1671. https://doi.org/10.1111/j.1469-8986.2012.01462.x

Seghier, M. L. (2008). Laterality index in functional MRI: methodological issues. *Magnetic Resonance Imaging*, *26*(5), 594-601. <Go to ISI>://000257272700002 (Not in File)

Seghier, M. L. (2019). Categorical laterality indices in fMRI: a parallel with classic similarity indices. *Brain Structure & Function*, *224*(3), 1377-1383. https://doi.org/10.1007/s00429-019-01833-9

Seghier, M. L., Kherif, F., Josse, G., & Price, C. J. (2011). Regional and Hemispheric Determinants of Language Laterality: Implications for Preoperative fMRI. *Human Brain Mapping*, *32*(10), 1602-1614. WOS:000295539600008 (Not in File)

Serrien, D. J., Ivry, R. B., & Swinnen, S. P. (2006). Dynamics of hemispheric specialization and integration in the context of motor control. *Nature Reviews Neuroscience*, *7*(2), 160-167. https://doi.org/10.1038/nrn1849

Shankweiler, D., & Studdert-Kennedy, M. (1975). A continuum of lateralization for speech perception. *Brain and language*, *2*(2), 212-225.

Sherwood, D. E. (2014). Aiming accuracy in preferred and non-preferred limbs: implications for programing models of motor control. *Frontiers in Psychology*, *5*. https://doi.org/ARTN 1236 10.3389/fpsyg.2014.01236

Smekal, V., Burt, D. M., Kentridge, R. W., & Hausmann, M. (2022). Emotion Lateralization in a Graduated Emotional Chimeric Face Task: An Online Study. *Neuropsychology*, *36*(5), 443-455. https://doi.org/10.1037/neu0000804

Smith, E. E., Reznik, S. J., Stewart, J. L., & Allen, J. J. B. (2017). Assessing and conceptualizing frontal EEG asymmetry: An updated primer on recording, processing, analyzing, and interpreting frontal alpha asymmetry. *International Journal of Psychophysiology*, *111*, 98-114. https://doi.org/10.1016/j.ijpsycho.2016.11.005

Speaks, C., Bauer, K., & Carlstrom, J. (1983). Peripheral Hearing LossImplications for Clinical Dichotic Listening Tests. *Journal of Speech and Hearing Disorders*, *48*(2), 135-139.

Speaks, C., Blecha, M., & Schilling, M. (1980). Contributions of monotic intelligibility to dichotic performance. *Ear and hearing*, *1*(5), 259-266.

Speaks, C., Niccum, N., & Carney, E. (1982). Statistical properties of responses to dichotic listening with CV nonsense syllables. *The Journal of the Acoustical Society of America*, *72*(4), 1185-1194.

Steenhuis, R. E., & Bryden, M. P. (1989). Different Dimensions of Hand Preference That Relate to Skilled and Unskilled Activities. *Cortex*, *25*(2), 289-304. https://doi.org/Doi 10.1016/S0010-9452(89)80044-9

Stroobant, N., Van Boxstael, J., & Vingerhoets, G. (2011). Language lateralization in children A functional transcranial Doppler reliability study. *Journal of Neurolinguistics*, *24*(1), 14-24. https://doi.org/10.1016/j.jneuroling.2010.07.003

Stroobant, N., & Vingerhoets, G. (2001). Test-retest reliability of functional transcranial Doppler ultrasonography. *Ultrasound in Medicine and Biology*, *27*(4), 509-514. https://doi.org/Doi 10.1016/S0301-5629(00)00325-2

Suarez, R. O., Golby, A., Whalen, S., Sato, S., Theodore, W. H., Kufta, C. V., Devinsky, O., Balish, M., & Bromfield, E. B. (2010). Contributions to singing ability by the posterior portion of the superior temporal gyrus of the non-language-dominant hemisphere: First evidence from subdural cortical stimulation, Wada testing, and fMRI. *Cortex*, *46*(3), 343-353. https://doi.org/10.1016/j.cortex.2009.04.010

Tenke, C. E., & Kayser, J. (2005). Reference-free quantification of EEG spectra: Combining current source density (CSD) and frequency principal components analysis (fPCA). *Clinical Neurophysiology*, *116*(12), 2826-2846. https://doi.org/10.1016/j.clinph.2005.08.007

Tenke, C. E., & Kayser, J. (2012). Generator localization by current source density (CSD): Implications of volume conduction and field closure at intracranial and scalp resolutions. *Clinical Neurophysiology*, *123*(12), 2328-2345. https://doi.org/10.1016/j.clinph.2012.06.005

Todor, J. I., & Doane, T. (1977). Handedness Classification - Preference Versus Proficiency. *Perceptual and Motor Skills*, *45*(3), 1041-1042. https://doi.org/DOI 10.2466/pms.1977.45.3f.1041

Tran, U. S., Stieger, S., & Voracek, M. (2014). Evidence for general right-, mixed-, and left-sidedness in self-reported handedness, footedness, eyedness, and earedness, and a primacy of footedness in a large-sample latent variable analysis. *Neuropsychologia*, *62*, 220-232. https://doi.org/10.1016/j.neuropsychologia.2014.07.027

Van der Haegen, L., Drieghe, D., & Brysbaert, M. (2010). The Split Fovea Theory and the Leicester critique: What do the data say? *Neuropsychologia*, *48*(1), 96-106. https://doi.org/10.1016/j.neuropsychologia.2009.08.014

Verhelst, H., Dhollander, T., Gerrits, R., & Vingerhoets, G. (2021). Fibre-specific laterality of white matter in left and right language dominant people. *Neuroimage*, *230*. https://doi.org/ARTN 117812 10.1016/j.neuroimage.2021.117812

Voyer, D. (1998). On the reliability and validity of noninvasive laterality measures. *Brain & Cognition*, *36*(2), 209-236.

Walker, R., & McSorley, E. (2006). The parallel programming of voluntary and reflexive saccades. *Vision Research*, *46*(13), 2082-2093. https://doi.org/10.1016/j.visres.2005.12.009

Wegrzyn, M., Mertens, M., Bien, C. G., Woermann, F. G., & Labudda, K. (2019). Quantifying the Confidence in fMRI-Based Language Lateralisation Through Laterality Index Deconstruction. *Front Neurol*, *10*, 655. https://doi.org/10.3389/fneur.2019.00655

Westerhausen, R. (2019). A primer on dichotic listening as a paradigm for the assessment of hemispheric asymmetry. *Laterality: Asymmetries of Body, Brain and Cognition*, *24*(6), 740-771.

Westerhausen, R., Passow, S., & Kompus, K. (2013). Reactive cognitive-control processes in free-report consonant–vowel dichotic listening. *Brain Cogn*, *83*(3), 288-296.

Westerhausen, R., & Samuelsen, F. (2020). An optimal dichotic-listening paradigm for the assessment of hemispheric dominance for speech processing. *PloS one*, *15*(6), e0234665.

Wexler, B. E. (1988). Dichotic presentation as a method for single hemisphere stimulation studies. In K. Hugdahl (Ed.), *Handbook of Dichotic Listening: Theory, methods, and research.* (pp. 85-115). Wiley & Sons

Wexler, B. E., & Halwes, T. (1983). Increasing the power of dichotic methods: The fused rhymed words test. *Neuropsychologia*, *21*(1), 59-66.

Wexler, B. E., & King, G. P. (1990). Within-modal and cross-modal consistency in the direction and magnitude of perceptual asymmetry. *Neuropsychologia*, *28*(1), 71-80.

Wilke, M., & Lidzba, K. (2007). LI-tool: A new toolbox to assess lateralization in functional MR-data. *Journal of Neuroscience Methods*, *163*(1), 128-136. WOS:000246934100017 (Not in File)

Wilke, M., & Schmithorst, V. J. (2006). A combined bootstrap/histogram analysis approach for computing a lateralization index from neuroimaging data. *Neuroimage*, *33*(2), 522-530. https://doi.org/10.1016/j.neuroimage.2006.07.010

Willems, R. M., Van der Haegen, L., Fisher, S. E., & Francks, C. (2014). On the other hand: including left-handers in cognitive neuroscience and neurogenetics. *Nature Reviews Neuroscience*, *15*(3), 193-201. https://doi.org/10.1038/nrn3679

Woodhead, Z. V. J., Rutherford, H. A., & Bishop, D. V. M. (2018). Measurement of language laterality using functional transcranial Doppler ultrasound: a comparison of different tasks. *Wellcome open research*, *3*, 104. https://doi.org/10.12688/wellcomeopenres.14720.2

Woodhead, Z. V. J., Thompson, P. A., Karlsson, E. M., & Bishop, D. V. M. (2021). An updated investigation of the multidimensional structure of language lateralization in left- and right-handed adults: a test-retest functional transcranial Doppler sonography study with six language tasks. *Royal Society Open Science*, *8*(2). https://doi.org/10.1098/rsos.200696

Xiang, L., Crow, T. J., Hopkins, W. D., & Roberts, N. (2020). Comparison of Surface Area and Cortical Thickness Asymmetry in the Human and Chimpanzee Brain. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhaa202

Young, A. W. (1982). Methodological theoretical bases. In J. G. Beaumont (Ed.), *Divided visual field studies of cerebral organisation* (pp. 11-27). Academic Press.

Supplementary Table 1: General description of the expert panel for each section in Round 1 and Round 2

	Round 1									**Round 2**								
	GEN	DLO	VHF	PER	PBR	EPR	fMRI	sMRI	fTCD	GEN	DLO	VHF	PER	PBR	EPR	fMRI	sMRI	fTCD
Number of experts (n)	102	27	28	46	42	20	28	26	13	95	25	26	40	46	16	26	21	13
Location (%)																		
- Europe	67	89	61	61	56	70	67	65	92	66	80	62	62	67	69	73	71	92
- North-America	21	7	25	26	29	25	19	15		17	12	15	25	20	19	8	10	
- Australasia	7	4	11	7	7	5	7	12	8	9	8	15	5	7	12	4	14	8
- Asia	3		4	4	2		4	8		5		4	2	2		15	5	
- South-America	3			2	5		4			3		4	5	4				
Female/male ratio (%)																		
- female	44	44	36	46	49	45	37	35	77	47	48	38	48	50	44	31	38	77
- male	54	56	64	52	51	55	63	62	15	53	52	62	52	50	56	69	62	23
- prefer not to say	2			2				4	8									
Academic field (%)																		
- cognitive neuroscience	58	63	71	54	51	70	78	77	62	59	60	65	55	50	62	77	76	62
- experimental psychology	29	30	29	39	39	20	7	4	31	31	32	35	40	41	25	8	5	31
- clinical (neuro)psychology	5	4		2	5	10	7	8	8	4	4		2	2	12	8	5	8
- medicine	5	4		4	2		4	8		3	4			2		4	5	
- mathematics/statistics	2				2		4			2			2	4		4	5	
- engineering	1							4		1							5	
Research experience (%)																		
- 1-5 years	3	4	7	4	2	5	4			3		4	5	4		4		
- 6-10 years	7	7	11	9	5	10		8	15	9	8	8	8	9	12	8	10	15
- 11-20 years	32	37	25	39	37	30	41	38	46	37	40	31	38	37	25	42	43	54
- 21-30 years	32	30	32	26	34	25	37	38	23	33	36	35	28	33	38	27	38	15
- more than 31 years	24	22	25	22	22	30	19	15	15	18	16	23	22	17	25	19	10	15
Number of publications (%)																		
- 6-10	5		4	7	5		4		15	5		4	8	7		4		15
- 11-30	20	19	18	24	24	25	7	8	23	22	16	15	28	26	19	12	14	23
- 31-60	25	37	32	30	24	25	22	23	38	27	40	31	20	26	25	19	24	38
- 61-100	20	15	18	9	22	15	30	19	8	20	16	23	22	22	25	31	24	8
- more than 100	30	30	29	30	25	35	37	50	15	26	28	27	22	20	31	35	38	15
Years since PhD (%)																		
- will obtain PhD within 1 year	2		4	2	2		4			2		4	2	2		4		
- 1-5 years	11	11	14	13	7	20	7	15	15	13	8	8	12	13	12	12	19	15
- 6-10 years	11	15	18	17	15	10	15	12	15	12	16	15	15	13	12	15	19	23
- 11-20 years	36	33	21	35	37	30	33	38	38	38	40	31	35	37	31	35	38	38
- 20-30 years	26	30	25	17	17	30	30	27	15	24	28	27	18	22	44	31	24	8
- more than 31 years	14	11	18	15	22	10	11	8	15	11	8	15	18	13		4		15

Supplementary Table 2: Statement ratings for the general section. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (🛇). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 95 experts (1 vote = 1.05%).

Statement ratings: General	☺☺	☺	😐	☹	☹☹	🛇	☺☺+☺
Laterality index relevance							
GEN#1. *We need a laterality index because it combines dependent measures from left and right sides into a single measure. Thus, it provides the convenience of representing an asymmetry directly, with a single score. For example, it allows the investigator to use a single score to represent laterality in a correlation matrix, an analysis of variance, or a multiple regression analysis.*	56.8	42.1	1.1	-	-	-	98.9
GEN#61. *Laterality indices are important, because there are cases in which the relative difference between the hemispheres/sides is more important than is the absolute level of independent left or right activity or performance. That is, no effect of higher left-sided scores may be expected if the right-sided scores are also higher. Consequently, relationships to other variables often are not observed if only absolute left and right activity/performance is examined and data of the left and right side are not related to each other by using an appropriate laterality index.*	30.5	51.6	13.7	2.1	-	2.1	82.1
GEN#2. *Guidelines on methods for quantifying and classifying laterality are needed because the findings and conclusions of a laterality study may vary according to the methods used to define and measure the phenomena of interest, and to analyse the obtained data. Science is best served when authors make well-reasoned choices of operational definition, method, and statistical analysis, and when they make available enough information to enable the reader to evaluate those choices.*	81.1	17.9	1.0	-	-	-	99.0
GEN#3. *Efforts should be made to make tests of behavioural and brain biases effective for testing across ages and species.*	27.4	50.5	14.7	5.3	-	2.1	77.9
GEN#4. *Quantifying laterality by considering the many factors and their mutual intertwinement that can influence laterality, is essential to understand when, why, and how laterality develops at the individual, population, and species levels.*	47.4	41.1	10.5	-	-	1.0	88.5
GEN#5rw. *Quantitative measures of lateralization are generally more useful than binary measures.*	50.5	32.6	11.6	2.1	1.0	2.1	83.1
GEN#62. *Binary or tertiary measures of lateralization may be appropriate (e.g. neurosurgical planning). When used the underlying quantitative measures (e.g. laterality indices) should also be given whenever possible.*	25.3	48.4	15.8	2.1	-	8.4	73.7
GEN#6. *During the execution of a given task, lateralization is dynamic and can change at different temporal and spatial scales.*	26.3	51.6	15.8	4.1	-	2.1	77.9

GEN#8. *For both fine and gross motor activities (e.g., throwing, kicking, drawing), tests of preference should be distinguished from tests of performance.*	54.7	33.7	5.3	-	-	6.3	88.4
GEN#9. *Ideally performance and preference scores should be measured and reported as both separate and composite LIs. This might be particularly relevant in relation to developmental studies and studies including participants with, or at risk for, functional disabilities.*	30.5	47.4	14.7	3.2	1.2	3.2	77.9
GEN#10. *An LI score may reveal a preference or a superior efficiency in a task or set of tasks. However, the non-preferred or less efficient limb may be preferred or more efficient for another task or set of tasks.*	45.3	40.0	9.5	1.1	-	4.2	85.3
GEN#11. *A statistic that tests whether there is significant bias to one side (z-test of proportions for accuracy; t-test for RT) is preferable to a conventional LI. The statistic provides a measure that is highly correlated to a conventional LI (L-R)/(L+R) but has the advantage that it can also identify individuals who are significantly lateralized (e.g., by using a p-value).*	11.6	26.3	35.8	16.8	1.1	7.4	37.9
Laterality index formulas							
GEN#12. *A laterality index that considers the left-right difference relative to the overall score is preferable to simple left minus right difference scores that are easily affected by the number of observations which makes them difficult to interpret. In other words, the denominator of the laterality index should be the sum of the observations, so that the difference score is presented as a proportion.*	55.8	33.7	9.5	1.0	-	-	89.5
GEN#13. *Agreement is needed on one single preferred version of the classical laterality index.*	29.5	40.0	17.9	8.4	4.2	-	69.5
GEN#14rw. *If you agreed with the previous statement, which of the following options would you prefer?*							
• *I don't favour consensus on the use of a classic laterality index.*		20.0					
• *Other.*		6.3					
• *The classic laterality index should always be (R-L)/((R+L)/2) regardless of the variable used. Using a denominator of (R+L)/2 ensures that the relative size of the effect is taken into account.*		7.4					
• *The classic laterality index should always be (R-L)/(R+L) regardless of the variable used.*		18.9					
• *The numerator of the classical laterality index should depend on the variable that is used. It is preferable to choose the numerator such that positive values indicate a rightward bias and negative values indicate a leftward bias. In this way plots are more intuitive and easier to interpret. The denominator should be the sum of both variables divided by two to ensure that the relative effect is taken into account. For example the formula for accuracy measures, where higher values denote higher performance, would be (R-L)/((R+L)/2). For reaction time measures, where higher values denote lower performance, the formula would be (L-R)/((L+R)/2).*		23.2					
• *The numerator of the classical laterality index should depend on the variable that is used. It is preferable to choose the numerator such that positive values indicate a rightward bias and negative values indicate a*		24.2					

leftward bias. In this way plots are more intuitive and easier to interpret. For example the formula for accuracy measures, where higher values denote higher performance, would be (R-L)/(R+L). For reaction time measures, where higher values denote lower performance, the formula would be (L-R)/(L+R).							
GEN#15. *The classic laterality index is the simplest way to consider the overall score and is sufficient in many contexts. However, its limitations and the use of more advanced indices such as phi (Repp, 1977) and lambda (Bryden & Sprott, 1981) should at least be considered.*	8.4	42.6	24.2	6.3	-	18.9	51.0
GEN#20. *We should come to a consensus regarding the scale of the laterality index.*	27.4	43.2	23.2	4.2	-	2.1	70.6
GEN#21rw. *If you agreed with the previous statement, which of the following options would you prefer?*							
• *I don't favour consensus on the use of a certain scale of the laterality index*		23.2					
• *Other*		2.0					
• *Scale the range of any laterality index between -1 and +1*		41.1					
• *Scale the range of any laterality index between -100 and +100. That is, multiply the laterality index by 100 and report as a percentage*		33.7					
Reporting laterality indices							
GEN#22. *Motivate the choice of the selected index. What characteristics make it particularly appropriate for these data?*	38.9	32.6	17.9	1.1	-	9.5	71.5
GEN#23. *Describe the way the laterality index was calculated. Present the formula and cite informative references.*	68.4	23.2	3.0	0	-	5.3	91.6
GEN#24. *Information is lost when raw scores are combined into an index. That information might be important for the purposes of the study. So, besides the LI, it is important to also report the R and L raw score values to further explore each side's contribution to the variation of interest. Preferably, this raw data is reported for each individual in supplementary material or in an open-access repository.*	45.3	33.7	13.7	4.2	1.1	2.1	79.0
GEN#25. *Always report standard error or 95% confidence interval around the LI.*	43.2	38.9	14.7	2.1	-	1.1	82.1
GEN#26. *Descriptive statistics for LI's should minimally include mean, median, standard deviation, inter-quartile range, minimum and maximum. This should be provided for each group separately.*	25.3	49.5	15.8	7.4	-	2.0	74.8
GEN#27. *For all LI group comparisons authors should indicate a measure of effect size in addition to test statistics as this is important for later meta-analyses and comparability between studies.*	48.4	40.0	8.4	-	1.1	2.1	88.4
GEN#28. *When reporting effect size, use Cohens' s original formula for the sake of simplicity and ease of comparison: d = (M1- M2)/ Sp, where M1 and M2 are means of right and left trials, and Sp is the pooled standard deviation.*	18.9	32.6	29.5	6.3	1.1	11.6	51.5

GEN#63. *There are two ways to calculate effect size for repeated measures: d-average (as defined in statement 28) and d-z-score (directly related to the test used, influenced by the correlation between the L and R scores). It would be better to always report both for meta-analyses.*	5.3	28.4	31.6	7.4	1.0	26.3	33.7
GEN#29. *Report individual laterality indices. Ideally, a graphic presentation of its distribution includes all individual data points (i.e., column scatter or violin plots).*	30.5	46.3	15.8	4.1	2.1	1.1	76.8
GEN#30. *Subjects' demographics (age, gender and handedness) should be provided in studies on laterality.*	71.6	26.3	1.1	-	1.1	-	97.9
GEN#31. *Researchers should routinely indicate whether the grand mean of a laterality index differs from zero (virtual symmetry).*	17.9	43.2	21.1	5.3	0.9	11.6	61.1
Reliability and validity							
GEN#33. *Make best efforts to provide test-retest reliability data for tasks that are not widely used.*	30.5	52.6	9.5	3.2	-	4.2	83.1
GEN#34. *Report internal reliability for laterality indices: split-half or Cronbach's alpha.*	20.0	38.9	23.2	4.2	1.1	12.6	58.9
GEN#35. *When reporting split-half reliability for laterality indices, the normality of the distribution must be considered to determine a suitable statistic (i.e. Pearson vs. Spearman).*	12.6	47.4	16.8	1.1	1.1	21.1	60.0
GEN#36. *The calculation of any index of laterality should be based on reliable estimates of R and L, therefore reliability of the measures used should also be reported.*	16.8	44.2	24.2	6.3	-	8.4	61.0
GEN#37. *A crucial step of any approach for a valid quantification of laterality is a careful selection of the items.*	38.9	44.2	8.4	1.1	0	7.4	83.1
GEN#38. *We should agree on the minimal number of occurrences to consider in order to calculate handedness indices.*	22.1	35.8	23.2	10.5	1.0	7.4	57.9
GEN#40. *If handedness is not included as an independent variable in the experimental design, laterality indices of (consistent and inconsistent) left- and right- handers should be included in the analysis. It is not advisable to include only participants who are strongly right-handed.*	16.8	33.7	21.1	9.5	14.7	4.2	50.5
Statistical concerns							
GEN#41. *The LI is likely to have statistical properties, especially reliability, that differ substantially from the statistical properties of the raw scores on which the index is based. The polarity and magnitude of the correlation between the index and another variable, such as overall performance, often will complicate interpretation of the data. Consequences of choosing a laterality ratio should be considered and explicated.*	9.5	47.4	21.1	4.2	2.1	15.8	56.9
GEN#45. *The unsigned (absolute) magnitude of a laterality index |LI| is of biological interest (degree of asymmetry in either direction). However, its population distribution often suffers from severe non-normality (floor at zero). Data analysis methods (e.g. for testing group differences) should be appropriate for this*.	21.1	53.7	12.6	2.1	1.1	9.5	74.8

GEN#50. *Data analyses should be based both on left and right side treated as a repeated measure factor to consider level of performance issues across experimental conditions, not only on a laterality index.*	16.8	26.3	26.3	13.7	2.2	14.7	43.1
GEN#51. *Laterality indices should not be compared or aggregated in meta-analyses across studies that used different computation methods and different population characteristics.*	22.1	27.4	21.1	16.8	6.3	6.3	49.5
Laterality index calibration and decomposition							
GEN#54rw. *Task performance should be reported and ultimately matched across groups and tasks.*	11.6	35.8	31.6	7.4	4.1	9.5	47.4
GEN#55rw. *Laterality research should be encouraged to decompose the chosen laterality index into two sub-components: direction (left/right) and absolute valence or strength (how much away from zero without +/- sign). This will lead to a more sophisticated interpretation of effects.*	21.1	44.2	20.0	7.4	1.0	6.3	65.3
GEN#56. *Use specific terminology for handedness; that is, use "hand preference" for the bias in spontaneous choice of hand for a given task; and use "hand skill difference" for the difference in skill between the hands in a given task, with handedness as the umbrella term.*	43.2	46.3	7.4	1.1	1.1	1.1	89.5
GEN#57. *For each measure, a unanimous categorization would be useful. For example, as regards handedness, a clear definition of right-handed, left-handed, and ambidextrous individuals should be reached.*	32.6	47.4	9.5	7.4	1.1	2.1	80.0
GEN#58. *When cut-off scores based on laterality indices are used to divide research participants into categories, or used as part of exclusion/inclusion criteria, the rationale for using a particular score should be articulated.*	57.9	37.9	4.2	-	-	-	95.8
GEN#64. *When cut-off scores based on laterality indices are used to divide research participants into categories, or used as part of exclusion/inclusion criteria, cut-off scores should be empirically validated and not arbitrary or only ad hoc.*	35.8	36.8	21.1	5.3	-	1.0	72.6
GEN#59. *We should refrain from using the term "normal" or "typical" lateralization pattern.*	15.8	27.4	28.4	18.9	5.3	9.5	43.2
GEN#60. *For interpretation, it is better to use left vs. right lateralization rather than typical vs. atypical lateralization, given the unknown size of lateralization bias for a given function and a particular population.*	32.6	36.8	17.9	8.4	-	4.3	69.4

Supplementary Table 3: Statement ratings for the dichotic listening technique. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (⊘). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 26 experts (1 vote = 3.84%).

Statement ratings: Dichotic listening technique	☺☺	☺	😐	☹	☹☹	⊘	☺☺+☺
Hearing deficits							
DL#1. *Exclude participants with hearing deficits in the speech frequency range of < 10 dB in either ear as tested by standard audiometry.*	26.9	30.8	15.4	7.7	11.5	7.7	57.7
DL#2. *Use a standard audiometry test to assess auditory threshold and interaural threshold differences.*	38.5	34.6	11.5	3.8	3.8	7.7	73.1
DL#3. *Laterality indices reported for dichotic listening should account for non-symmetrical (between right and left ear) hearing acuity (i.e., exclude participants, use threshold asymmetry).*	34.6	50.0	7.7	-	3.8	3.8	84.6
DL#4. *For dichotic listening tests, we need to include a check to ensure headphones are the right way around.*	77.0	23.0	-	-	-	-	100
Reliability							
DL#5. *The number of trials should be enough to ensure an acceptable degree of retest reliability*.	61.5	30.8	7.7	-	-	-	92.3
DL#6. *State the retest reliability of the index scores. The index scores may be unreliable even if the left-ear and right-ear scores are reliable.*	15.4	34.6	30.8	7.7	-	11.5	50.0
Task construction							
DL#7. *The spectral and temporal overlap of dichotic stimuli across channels should be maximized to promote cross-channel competition and stimulus fusion.*	38.5	26.9	23.1	3.8	-	7.7	65.4
DL#8. *Single dichotic stimulus presentation per trial is optimal to assess hemispheric lateralization as it reduces the working memory load compared to multi stimulus paradigms.*	19.2	34.6	19.2	15.4	11.6	7.7	53.8
DL#9rw. *Stimuli used need to be appropriate for the sample. This should be tested by additionally introducing trials in which the same stimulus is presented to both ears.*	30.8	42.3	19.2	3.8	-	3.7	73.1
DL#11. *Intertrial effects, like negative priming, need to be considered when designing the paradigm.*	11.5	38.5	26.6	7.7	-	15.4	50.0
DL#12. *Dichotic stimuli might at first appear confusing to the participant. Test trials before the proper experiment might thus help the participant to familiarize with the testing situation.*	50.0	50.0	-	-	-	-	100
Data reporting							

DL#13. *Use a laterality index that standardizes the interaural difference to the overall level of performance (sum of left- and right-ear correct recall).*	15.4	53.8	15.4	3.9	-	11.5	69.2
DL#14. *Report left- and right-ear correct recall in addition to the laterality index to identify which side contributes to changes/differences in laterality.*	57.7	34.6	-	-	-	7.7	92.3
DL#15. *The laterality index should be determined so that left-ear preference results in a negative value, and a right-ear preference in a positive value.*	34.6	50.0	11.5	-	-	3.9	84.6
DL#17. *Various measures can be collected in dichotic listening to assess perceptual laterality (e.g., the score for each ear might be number of correctly identified stimuli, reaction times, signal-detection sensitivity, disruption from noise or interference, delayed recall, etc.), laterality indices need to consider difference in the characteristics of these measures.*	23.1	61.5	11.5	3.9	-	-	84.6
DL#18. *Multiple dependent variables, e.g., magnitude of the right-ear advantage and proportion of subjects who show a right-ear advantage, should be calculated.*	19.2	57.7	19.2	3.9	-	-	76.9

Supplementary Table 4: Statement ratings for the visual half-field technique. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (⊘). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 26 experts (1 vote = 3.84%).

Statement ratings: Visual half-field technique	☺☺	☺	😐	☹	☹☹	⊘	☺☺+☺
Eye movements							
VHF#1. *Eye tracking is good but not necessary, if other measures are taken to make sure participants fixate well (e.g., if participants are required to process information on some trials at the fixation location.*	46.2	46.2	3.8	3.8	-	-	92.4
VHF#2. *To avoid short saccades (<100 ms), the fixation stimulus must remain visible while the parafoveal stimuli are presented.*	38.5	38.5	15.4	7.6	-	-	77.0
Task construction							
VHF#3rw. *If manual responses are measured, stimulus-response compatibility effects should be avoided (e.g., by switching hand during the experiment or requiring bimanual responses).*	54.0	35.0	12.0	-	-	-	89.0
VHF#5. *Bilateral presentation (i.e., simultaneous presentation of two stimuli in LVF and RVF) is better than unilateral presentation.*	26.9	11.5	26.9	19.2	7.7	7.7	38.4
VHF#6. VHF presentations of stimuli should be in the range of 100-180 ms (depending on task difficulty).	34.6	42.3	15.4	3.8	-	3.8	76.9
VHF#7. *In VHF studies with bilateral presentation and an arrow in the middle pointing to the target stimulus, presentation times up to 200 ms are possible.*	15.4	34.6	26.9	3.9	-	19.2	50.0
VHF#8. *Stimulus eccentricities should be 1 degree visual angle off fixation, and not exceed 6 degrees visual angle.*	23.1	42.3	15.4	7.7	3.8	7.7	65.4
Reliability							
VHF#9rw. *Researchers should aim for 100+ observations per condition, in particular when RT is your dependent variable.*	7.7	38.5	11.5	34.6	3.8	3.8	46.2
VHF#10. *VHF studies should use a chin rest to ensure that head movements are minimised and constant distance from the monitor is maintained (i.e., faster responses with the left hand to LVF stimuli and faster responses with the right hand to RVF stimuli).*	34.6	34.6	11.5	11.5	3.9	3.9	69.2

VHF#12. *To investigate individual differences in VHF differences, researchers must calculate the reliability of the VHF differences (e.g., by having two blocks of trials and calculating the correlation between the VHF differences of both blocks and use the Spearman-Brown formula).*	3.9	30.8	34.6	19.2	-	11.5	34.7
Analysis pipeline							
VHF#13. *A laterality index should take overall level of performance into account, not only the left-right difference.*	26.9	57.7	11.5	0	0	3.8	84.6
VHF#14. *In VHF studies, researchers should report laterality indices for both response time and accuracy measures, if possible.*	50.0	42.3	7.7	-	-	-	92.3
VHF#15. The calculation of response time-based laterality indices should include correct responses only.	65.4	26.9	3.8	3.8	-	-	92.3
Data reporting							
VHF#17. *In VHF studies, researchers should report LVF and RVF performances (means and standard deviations/errors) in addition to laterality indices, because the calculation of laterality indices results in a loss of information*	65.4	23.1	-	7.7	-	3.8	88.5

Supplementary Table 5: Statement ratings for performance asymmetries. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (🛇). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 40 experts (1 vote = 2.5%).

Statement ratings: Performance asymmetries	☺☺	☺	😐	☹	☹☹	🛇	☺☺+☺
Defining laterality							
PA#1rw. *For motor tasks involving objects, the spatial location of the object relative to each hand should be clearly defined (e.g., in terms of a central body point from which deviations are measured; object placed at equal distances from each limb, etc).*	72.0	28.0	-	-	-	-	100
PA#3. *Definition of peri-personal and extra-personal space should be clarified.*	22.0	44.0	33.0	-	-	-	66.0
PA#4. *Methods need to ensure there are no implicit and explicit methodological or perceptual biases caused by apparatus positioning.*	50.0	39.0	11.0	-	-	-	89.0
Paradigm construction							
PA#5rw. *We should decide whether a general laterality index of performance should include bimanual activities or not.*	27.8	44.4	16.7	5.6	5.6	-	72.2
PA#5bis. *If you agreed with the previous statement what would be your recommendation?*							
• Exclude bimanual activities		44.4					
• I did not agree with the previous statement		16.7					
• Include bimanual activities, in which case the laterality index will need to be adapted, e.g. R-L/R+L+Bm		16.7					
• Other		22.2					
PA#7rw. *In research that includes bimanual activities in the laterality index, we need to distinguish between symmetrical and asymmetrical activities (and specify whether the asymmetry involves a functional and/or structural dominance).*	38.9	33.3	16.7	-	5.6	5.6	72.2
PA#8. *Where tasks can be performed from left to right (e.g., Tower of London, Tower of Hanoi), the starting direction should be counterbalanced across studies.*	50.0	33.3	5.6	5.6	-	5.6	83.3
PA#9. *For performance tests on both limbs and where the subject has identified one limb as preferred, the preferred limb should always be tested first.*	11.1	5.6	16.7	44.4	16.7	5.6	16.7
Relevant background information							

PA#11. *Researchers should take note of the possible effects of prior experience on current biases, e.g., keyboard/piano lessons for handedness, soccer training for footedness.*	44.4	27.8	22.2	-	-	5.6	72.2
Use of different behavioural measures							
PA#12. *Tests of motor preferences should reflect validated measures that have a known empirical basis and should be based on multiple behaviours (both frequently and infrequently performed).*	22.0	39.0	22.0	6.0	6.0	6.0	61.0
PA#13. *For a motor task, always assess the direction and degree of lateralization. Therefore, for preference, assess choice of limb, i.e., direction, whether the task was performed by the right or the left limb; for performance, i.e., degree of difference, assess how well the task was performed by such measures as the number of errors, time before responding, and after responding, the speed of performance and the smoothness of hand/arm movement.*	27.8	44.4	16.7	11.1	-	-	72.2
PA#14rw. *In research that includes bimanual activities in the laterality index, we need objective markers for distinguishing them from unimanual activities.*	39.0	39.0	11.0	-	-	11.0	78.0
PA#16rw. *Performance tests should place reasonable constraints on time and task, and here, as elsewhere, "reasonable" means taking the age and capacities of the participant into account. These constraints should be clearly defined.*	5.6	61.1	27.8	-	-	5.6	66.7
PA#17. *Assess both direction and degree of lateralization of lower and upper limbs.*	22.2	33.3	22.2	5.6	5.6	11.1	55.5
PA#18. *Motor performance asymmetry is task-specific and should not be expressed through averages across tasks.*	22.2	44.4	16.7	11.1	5.6	-	66.6
Use of kinematic analysis							
PA#22. *Measures of functional performances with each hand should be as detailed as possible, e.g., at the level of kinematics.*	11.1	16.7	33.3	27.8	5.6	5.6	27.7
PA#23. *When hand performance is measured through kinematic recording/analysis, markers should be attached at comparable anatomical landmarks on each hand and across studies (e.g., wrist marker at Lister's tubercle). State the reasons for choosing those landmarks.*	44.4	33.3	16.7	-	-	5.6	77.7
Errors, validity and reliability							
PA#25. *In tests with a continuous movement parameter, such as copying a line or a spiral, the result is the achieved time plus penalization converted to time.*	5.6	11.1	33.3	22.2	-	27.8	16.7
PA#38. *Where possible, performance tests should be videotaped to allow for estimating performance reliability.*	27.8	33.3	22.2	16.7	-	-	61.1
PA#39. *Motor performance asymmetry should be based on averages of at least three trials for each limb.*	27.8	61.1	5.6	-	-	5.6	88.9
PA#40. *The number of trials should be enough to ensure an acceptable degree of retest reliability.*	33.0	56.0	11.0	-	-	-	89.0

PA#41rw. *Quantification of limb performance should be based on at least four trials.*	11.1	55.6	16.7	5.6	-	11.1	66.7
PA#43. *The nature of the motor movement considered should be explicit and homogenous to calculate handedness indices (i.e., the function or intention of the participant, depending on the task or activity focused on).*	16.7	38.9	22.2	5.6	-	16.7	55.6
PA#44. *The task and the dependent variables should be selected to avoid floor and ceiling effects.*	38.9	50.0	5.6	-	-	5.6	88.9
Laterality index formulas							
PA#46. *The correlation between left-side and right-side scores should be computed. A strong correlation implies that the retest reliability of the difference score, and of indices derived from the difference score, will be low in retest reliability.*	5.6	27.8	38.9	-	-	27.8	33.4
PA#47. *Ideally, multiple dependent variables, e.g., magnitude of the performance asymmetry and proportion of subjects who show the asymmetry, should be calculated.*	5.6	61.1	33.3	-	-	-	66.7
PA#48. *Use hypothesis test criteria for classifying a participant as right-/left-handed (latent class analysis, confidence interval, etc.), rather than a selected cut-point (e.g., a value from formula (R-L)/(R+L)).*	5.6	44.4	33.3	11.1	5.6	-	50.0
PA#49. *The percentage performance difference between the preferred and non-preferred body side should be reported, providing a clearer measure of how much the sides are asymmetric.*	5.6	61.1	11.1	16.7	-	5.6	66.7
PA#50. *Where a cut-off point is used to categorize behavioural laterality, it should be clearly stated (e.g., consistent footedness is defined as an LI of ± 80 or above; mixed footedness, as an LI between -79 and +79).*	56.0	44.0	-	-	-	-	100
PA#51. *In performance tests, where the result is time and we don't know the participant's preference, the LI is calculated according to the formula LQ= (R/R+L)*100, when R is the performance of the right limb, L is the performance of the left limb.*	5.6	22.2	50.0	11.1	5.6	5.6	27.8
PA#52. *In performance tests, where the result is time and we already know the participant's preference, the LI is calculated according to the formula LQ=(P/P+NP)*100 , when P is the performance of the preferred limb, NP is the performance of the non-preferred limb.*	11.0	28.0	28.0	11.0	11.0	11.0	39.0
PA#53. *In performance tests, where the result is the number of successful attempts (e.g., pegboard test) and we don't know the participant's preference, the LI is calculated according to the formula LQ=(R/R-L)*100 , when R is the performance of the right limb, L is the performance of the left limb.*	6.0	33.0	33.0	17.0	6.0	6.0	39.0
PA#54. *In performance tests, where the result is the number of successful attempts (e.g., pegboard test) and we already know the participant's preference, the LI is calculated according to the formula LQ=(P/P-NP)*100, when P is the performance of the preferred limb, NP is the performance of the non-preferred limb.*	11.1	27.8	27.8	16.7	11.1	5.6	38.9
Comprehensive data reporting							

PA#55. *For assessing asymmetries in manual performance, mean values as well as variability should be reported.*	44.0	56.0	-	-	-	-	100
PA#56. *Given that performance differences between the hands are influenced by the task completed, researchers should include averaged raw data by hand and sex in an appendix.*	5.6	44.4	27.8	11.1	11.1	-	50.0
Performance asymmetry in children							
PA#58. *We will need to agree on a specific set of guidelines that acknowledge the special dynamics associated with performance asymmetry in infants and children.*	22.2	61.1	16.7	-	-	-	83.3
PA#26. *To assess handedness in children, no fewer than 15 trials should be the standard minimum number.*	11.1	27.8	27.8	16.7	16.7	-	38.9
PA#27rw. *To assess the left-right direction of lateralization: a. for upper limbs, a middle line crossing task must be used; for young children (perhaps from 3 to 7 or 8), a drawing task should be used instead of writing. For testing which hand is used for throwing a ball at a target, always take into account the child's age (pre-school, school); b. for lower limbs, e.g., kicking a tennis ball at a target repeatedly.*	11.1	38.9	11.1	27.8	5.6	5.6	50.0
PA#28. *In tests with a discontinuous movement parameter (e.g., pegboard tests, dot-filling tests) where milestones of motor development are considered (i.e., adequate difficulty of the test), dot filling and copying line are not recommended for pre-school children; for them, pegboard tests are more appropriate.*	5.6	55.6	16.7	5.6	-	16.7	61.2
PA#29. *Unimanual versus Bimanual reaches and grasps should be defined. Strictly speaking, a unimanual reach means that only one hand is used for reaching and grasping; a bimanual reach means that both hands are used. Depending on the child's age, there may be no discernible difference between the hands, either for starting to reach (initiating movement toward the object) or for finishing (grasping the object). Eventually, one hand will take the lead by reaching first and/or by grasping first. Alternatively, the hand that starts second may be the first hand to grasp. A maximum delay between movements must be stipulated for a reach to be considered bimanual. If the second hand does not begin to move until the first hand has grasped the object, this should not be considered a bimanual reach or grasp.*	16.7	61.1	11.1	-	-	11.1	77.8
PA#30. *When calculating the LI, clearly explain whether it is calculated from the hand used for reaching (approach phase) or from the hand used for grasping (picking-up phase). In infants, reaching and grasping do not always yield identical LIs. Indeed, it is not unusual for infants to start reaching for an object with one hand only to grasp it with the other hand.*	28.0	61.0	-	-	-	11.0	89.0
PA#31. *LI should be calculated using objects that do not afford bimanual manipulation. When objects afford bimanual manipulation, it is more difficult to infer the child's actual intent, i.e., whether the child intended to grasp the object with one hand with the goal of using the other hand for manipulation or to do the reverse.*	16.7	61.1	11.1	5.6	-	5.6	77.8

PA#33. *Bimanual reaches/grasps are an important part of most infants' behavioural repertoire. Not including bimanual grasps underestimates the number of non-lateralized infants. If researchers decide to remove bimanual reaches/grasps from their LI formula, they should state the number that were removed.*	27.8	44.4	11.1	5.6	5.6	5.6	72.2
PA#34. *Limb preference indices of infants and children should be based on frequency of limb use in experimental or natural settings.*	22.2	66.7	5.6	5.6	-	-	88.9
PA#35. *Parental reports of children's hand and arm preference are not adequate (neither necessarily valid nor reliable) for assessing laterality. Instead, wherever possible, use performance-based measures of preference (e.g., which hand is more likely to cross the midline in a test of reaching); they are more likely to be valid and reliable.*	44.4	44.4	5.6	5.6	-	-	88.8
PA#36. *Infants may refuse to continue the test at some point for many reasons, leading to variable trial numbers across infants of the same study or even across studies. When this happens, the LI formula should not try to compensate for these variations in data by dividing the difference between RH and LH grasps by the square root of the total number of trials. When using this formula, with a similar ratio between N of RH grasps and N of LH grasps, the more trials the infants have, the higher is their LI and the more lateralized they appear. There is no evidence indicating that a stronger LI obtained with more trials is a more accurate index of an infant's handedness than a LI based on fewer trials.*	6.0	28.0	17.0	17.0	6.0	28.0	34.0

Supplementary Table 6: Statement ratings for preference bias reports. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (🛇). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 45 experts (1 vote = 2.22%).

Statement ratings: Preference bias reports	☺☺	☺	😐	☹	☹☹	🛇	☺☺+☺
Setting the standards							
PBR#1. *A consensus should be reached determining a gold-standard for assessment of hand preference.*	35.6	42.2	13.3	6.7	2.2	-	77.8
PBR#2. *If you agreed with the previous statement, which of the following options would you prefer for determining hand preference?*							
• *Edinburgh Handedness Inventory (Oldfield, 1971)*		37.8					
• *I do not favour consensus on a gold standard assessment of hand preference*		22.2					
• *Inventory of Global Lateral Preference (Marim, 2011)*		8.9					
• *Other*		20.0					
• *Waterloo Handedness Questionnaire (Steenhuis & Bryden, 1989)*		11.1					
PBR#3. *Handedness should be treated as a categorical variable.*	6.7	17.8	26.7	28.9	17.8	2.2	24.5
PBR#4. *Instead of LI, consider the use of a proportion of hand use (such as, R/(R+L)).*							
PBR#5. *A complete laterality assessment should include both i) a measure of preference (e.g., Edinburgh Handedness Inventory) and ii) a direct assessment of relative skill (e.g., pegboard task).*	17.8	37.8	20.0	20.0	2.2	2.2	55.6
PBR#6rw. *Inventory items should be limited to tasks that are common across a broad range of cultures and should be translated into multiple languages; culturally biased behaviours (e.g., feeding, grooming, or social greeting actions), where included, should be flagged for exclusion when inappropriate.*	22.2	55.6	17.8	4.4	-	-	77.8
PBR#7. *Clear analysis standards/procedures should be developed and agreed upon to create individual and population-level laterality profiles that include assessments of both sensory (visual, auditory) and motor (handedness, footedness) biases.*	22.2	51.1	17.8	4.4	-	4.4	73.3
Definitions							
PBR#8. *A standard glossary of terms should be developed to support meta-analyses and systematic reviews. Terms requiring definition: handedness, footedness, eyedness, earedness, preference (hand/foot/eye/ear), lateral preference, motor laterality degree/degree of laterality, laterality indices.*	40.0	48.9	8.9	2.2	-	-	88.9

PBR#9. *All inventory items should be fully described; if inventory items are part of an established, validated questionnaire (e.g., Edinburgh Handedness Inventory), authors should provide a reference to where a complete description may be found.*	62.2	33.3	4.5	-	-	-	95.5
PBR#10. *Classification labels (i.e., weak, moderate, or strong left- or right-handedness; mixed-handedness, ambidexterity) require clear and consistent definitions. Definitions should include scoring criteria (i.e., applicable ranges) from common/standard assessments.*	62.2	31.1	4.4	2.2	-	-	93.3
Test construction							
PBR#11. *Hand Preference/Skill is not universal; one hand may be dominant for a given task (or set of tasks), while the other may be dominant for another task (or set of tasks). Questionnaires should therefore measure preference across multiple criteria. Suggested criteria include fine motor skills (e.g., writing), gross motor skills (e.g., swing a bat), open/cyclic/continuous actions (e.g., stirring a pot), closed/discrete actions (e.g., reaching, grasping), ballistic actions (e.g., throwing), and communicative gestures (e.g., waving, pointing).*	31.1	53.3	8.9	4.4	2.2	-	84.4
PBR#12. *Inventory items should have a fixed number of response options.*	31.1	40.0	15.6	11.1	-	2.2	71.1
PBR#13. *If you agree with the previous statement, which of the following options would you prefer:*							
• *5 response options, e.g., "always left, usually left, equal/no preference, mostly right, always right," so that mixed/non-binary patterns of handedness may be detected.*		53.0					
• *A simple choice between three response categories (RH, LH, Either) is more reliable than using a gradual response scale (always RH, most of the time RH, etc.).*		16.0					
• *I do not favour consensus on a fixed number of response options.*		27.0					
• *Other*		4.0					
PBR#14. *Laterality is a multivariate construct; therefore, laterality should be measured using a multiple-item questionnaire to allow assessment of degree, as well as direction, of preference.*	51.1	37.8	8.9	-	-	2.2	88.9
PBR#16rw. *We should decide whether a general laterality index of preference must include bimanual activities or not.*	35.6	44.4	15.6	4.4	-	-	80.0
PBR#16bis. *If you agreed with the previous statement what would be your recommendation:*							
• *Exclude bimanual activities.*		36.0					
• *I did not agree with the previous statement.*		18.0					
• *Include bimanual activities.*		47.0					
Laterality index formulas							
PBR#17. *Consensus is needed on how to handle 'both/either' in quantifying hand preference.*	33.3	42.2	17.8	6.7	-	-	75.5
PBR#17bis. *If you agreed with the previous statement, what solution would you prefer?*							

• *“Either hand” responses should be excluded from the calculation of a laterality index, such that only clearly lateralized unimanual actions are considered (resulting in fewer ambiguous LI outcomes).*		13.3					
• *“Either hand” should be added to the denominator in a laterality index, such that bimanual or non-lateralized actions lower the calculated laterality index.*		64.4					
• *I did not agree with the previous statement.*		15.6					
• *Other*		6.7					
PBR#18. *If bimanual activities are probed in the laterality index, we need to distinguish between symmetrical and asymmetrical activities (and specify whether the asymmetry involves a functional and/structural dominance).*	24.4	40.0	6.7	11.1	-	17.8	64.4
PBR#19. *If the test includes bimanual items, each item should define the hand used for the active part of the action (for instance for scissors, the hand holding the scissors for cutting should be the one reported, not the one holding the sheet of paper, etc.).*	53.3	40.0	4.4	-	-	2.3	93.3
PBR#20. *Response paradigms should involve respondents providing one simple answer per question.*	20.0	42.2	17.8	2.2	-	17.8	62.2
Reliability and validity							
PBR#23. *Every inventory or questionnaire should provide its psychometric validation before being used for a specific purpose. Any variations/modifications should be clearly labelled and described.*	40.0	42.2	11.1	2.2	-	4.4	82.2
Infants and young children							
PBR#24. *Self-report and parental report via survey is not a sufficient indicator of manual biases - especially in the case of children. With younger children, a test with real objects (‘show me how you usually use this pencil, etc.’) is better than a pretend execution of the action without objects (‘show me with your hands how you use a pencil, etc’) and even better than a questionnaire (‘with which hand do you?’).*	37.8	44.4	4.4	4.4	-	8.9	82.2
PBR#25. *With children one should stick to RH, LH, or Either hand.*	13.3	22.2	35.6	11.1	6.7	11.1	35.5
Scoring and classification							
PBR#28. *The term ambidextrous should be used to indicate no preference between left and right for one specific task (e.g., individual can write with both left and right hand) and the term mixed handedness should be used to indicate preference of different hands for different tasks (e.g., use the right hand for some tasks and the left for others). In the case of laterality indices, we should avoid using the term ambidextrous and mixed-handedness and rather refer to scores in a range around zero representing equal performance between the two hands on a specific task.*	35.6	24.4	26.7	8.9	-	4.4	60.0
PBR#29. *Ambidexterity should refer to equality of performance, rather than ambiguity in preference. In other words, ambidexterity denotes an equality of performance ability, regardless of (daily) preference. If one writes equally well with both hands, but prefers one hand most of the time, then this person is ambidextrous. Preference reports should make this distinction.*	17.8	46.7	17.8	11.1	2.2	4.4	64.5

PBR#30. *We should consider multicollinearity of used items in measures of hand preference. Probing the hand preference for 7-8 tool items, still probes only the ‘tool’ construct of hand preference, while other factors of laterality preference remain unexplored.*	4.4	51.1	20.0	4.4	2.2	17.8	55.5

Supplementary Table 7: Statement ratings for electrophysiological recording. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (🛇). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 17 experts (1 vote = 5.88%).

Statement ratings: Electrophysiological recording	☺☺	☺	😐	☹	☹☹	🛇	☺☺+☺
Recording standards							
EPR#1. *The arrangement of the experimental setup should be as symmetrical as possible with respect to the participant's midline*.	58.8	35.3	5.9	-	-	-	94.1
EPR#2. *The EEG signal should be of comparable quality (e.g., SNR, impedance) across the two hemispheres.*	82.0	18.0	-	-	-	-	100
EPR#3. *Ensure that the recorded activity is not a result of lateral eye movements.*	76.0	24.0	-	-	-	-	100
EPR#16. *In order to enhance the usability of EEG and neuroimaging laterality data in cooperative research networks, they should be recorded following the FAIR principles and should use the BIDS EEG / BIDS neuroimaging data organization scheme.*	17.6	17.6	35.3	5.9	-	23.5	35.2
Reference schemes							
EPR#15. *The choice for a specific EEG reference needs to be clearly stated and argued, and its implications for data laterality analysis need to be discussed.*	64.7	29.4	5.9	-	-	-	94.1
EPR#4. *EEG asymmetry indices should be based on data that has been current-source density transformed to provide more precise estimates of local laterality.*	29.4	29.4	17.6	17.6	5.9	-	58.8
EPR#5. *Preferably, reference-free EEG analysis methods at sensor and brain level should be considered and used*.	23.5	35.3	23.5	11.8	5.9	-	58.8
EPR#17. *Meta-analyses of EEG asymmetry should include the reference montage as a factor that might explain the effect size.*	58.8	29.4	5.9	-	-	5.9	88.2
Calculating and reporting asymmetries							
EPR#6. *EEG asymmetry should be based on at least 100 artifact-free epochs.*	11.8	29.4	5.9	35.3	17.6	-	41.2
EPR#7. *Lateralization of EEG activity should be computed between homologous (groups of) pairs of electrodes across the two hemispheres.*	41.0	59.0	-	-	-	-	100
EPR#8rw. *Frontal EEG asymmetry should be reported as ln(right) - ln(left) alpha activity (8 -13 Hz), with higher scores putatively indexing relatively greater left frontal activity, and lower scores indexing relatively less left frontal activity*.	29.4	35.3	17.6	11.8	-	5.9	64.7

EPR#9. *It is also possible to use the laterality coefficient (LC) computed as LC = (R-L)/(R+L), where R denotes alpha power at the right hemispheric electrode position and L denotes alpha power at the homologous left hemisphere position. For mathematical reasons, in the small physiologically expectable range of relative differences between the EEG alpha power at two homologous electrodes, the correlation between LC and the metric (lnR - lnL) is very close to 1. Compared to the metric (lnR - lnL), the range of LC is confined to -1 to +1, which makes the meaning of scores intuitive; and as LC is a relative score, hemispheric differences are easily comparable between different electrode positions, conditions, and studies. LC is also commonly used in other fields of laterality research.*	17.6	47.1	17.6	5.9	-	11.8	64.7
EPR#18. *For studies of individual differences in EEG laterality, the reliability of the laterality index should be reported (e.g., split-half or test-retest).*	17.6	58.8	17.6	-	-	5.9	76.4
EPR#14. *EEG asymmetries should not be phrased as 'higher' or 'lower' symmetry, but the phrasing should always include the direction of asymmetry (e.g., greater relative right frontal activity, greater relative left hemisphere activation.).*	52.9	47.1	-	-	-	-	100
Data analysis and reporting							
EPR#10. *Data-driven analytic approaches (multivariate data decomposition) are preferable over post-hoc and a priori defined measures.*	11.8	5.9	29.4	47.1	5.9	-	17.7
EPR#11. *Besides laterality indexes (or laterality tests in MANOVA), effects in each hemisphere should be reported.*	52.9	29.4	5.9	5.9	5.9	-	82.3
EPR#12. *Significant correlations with laterality indices should be followed up with correlations on each side.*	23.5	58.8	5.9	5.9	5.9	-	82.3
EPR#13. *If laterality (effects) is (are) predicted in specific areas, they should always be compared to laterality (effects) in other control areas.*	17.6	52.9	23.5	0	5.9	-	70.5

Supplementary Table 8: Statement ratings for functional task-related MRI. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (🛇). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 26 experts (1 vote = 3.84%).

Statement ratings: Functional task-related MRI	☺☺	☺	😐	☹	☹☹	🛇	☺☺+☺
Tasks and paradigms							
fMRI#2. *We should decide on a consensus for the optimal control condition when determining activation LQ's from functional task-related fMRI (as rest vs. different forms of active control can affect the extent of lateralization).*	26.9	46.2	3.8	7.7	3.8	11.5	73.1
fMRI#3. *We need to establish a task-related MRI 'hand preference paradigm' whose laterality index can be compared with asymmetries in hand preference/performance measures and whose LI can be used in ways comparable to those of other task-related MRI localizers.*	7.7	57.7	11.5	7.7	-	15.4	65.4
fMRI#4. *Language dominance assessment: use several tasks instead of only one and express lateralization in terms as LI for production vs. LI for comprehension; LI for semantic, LI for phonological, etc.*	34.6	46.2	7.7	3.8	-	7.7	80.8
Reliability							
fMRI#5. *Clarify the issue of bilateral representation. In general, subjects cannot be reproducibly categorized as bilateral. Recommendation to classify a subject as 'bilateral' only based on the use of various methods of calculation; adopt a 'criterion of stability'.*	11.5	42.3	26.9	7.7	-	11.5	53.8
fMRI#6. *We need to establish test-retest reliability of laterality indices for a number of standard task-related fMRI paradigms.*	42.3	53.8	-	-	-	3.8	96.1
fMRI#7. *Reproducibility of laterality indices is bounded by the intrinsic reproducibility of fMRI across sessions.*	23	42	12	12	-	12	65
fMRI#9. *When calculating lateralization indices, the quality of the underlying data should routinely be assessed to avoid issues of data scarcity.*	34.6	57.7	3.8	-	-	3.8	92.3
fMRI#10. *In clinical care, laterality index-based regression equations predicting neurosurgical risk to cognition should only be used if the protocol they are based on is precisely replicated.*	34.6	42.3	11.5	-	-	11.5	76.9
Region-of-interest							
fMRI#11. *All voxel-based measures of lateralization should only be performed after spatial normalization to a symmetric template.*	19.2	50	11.5	3.8	3.8	11.5	69.2
fMRI#12. *Functional hemispheric differences can be determined for predefined region-of-interest or on voxel-/vertex-level.*	23	65	-	-	-	12	88

fMRI#15. *When calculating the laterality index based on the number of significantly activated voxels (voxel count) in two homological regions, we need to account for the regions' size*	42.3	46.2	3.8	3.8	-	3.8	88.5
fMRI#16*. The regions of interest and reference brain used to calculate laterality indices should be publicly available with documentation on their source and construction*	53.8	38.5	-	-	-	7.7	92.3
fMRI#17. *Calculating laterality indices over geometrically homologous regions may not inform about true lateralization of a cognitive function, if one of the two regions has no competence in said function.*	19	35	3.8	19	3.8	19	54
fMRI#20. *When signal magnitude methods are used, particular attention should be paid to: (a) delineation of ROI: ROIs should be small and comparable between hemispheres as homologous, (b) to avoid miscalculation due to deactivated regions, a solution is to only compare BOLD intensity in those voxels that are most strongly activated within each ROI.*	15.4	23.1	30.8	7.7	3.8	19.2	38.5
fMRI#21. *We need to establish consensus on how we determine the region of interest for calculating a laterality index for standard task-related MRI paradigms.*	23.1	46.2	15.4	7.7	-	7.7	69.3
fMRI#22. *We should use regions of interest with comparable sizes in left and right hemispheres*.	19.2	42.3	23.1	7.7	-	7.7	61.5
fMRI#23. *It is not adequate to claim that the activation is lateralized by observing activation cluster in one hemisphere if you have not actually tested whether it's significantly stronger than in homologous voxels in the other hemisphere.*	38.5	38.5	7.7	7.7	-	7.7	77
Analysis (Method)							
fMRI#25. *To calculate the LI, two measures exist to assess LH and RH activity: signal extent and signal magnitude. Signal extent refers to the absolute number of voxels that shows activity over certain threshold in each hemisphere. Although both measures yield to similar LI and curves, signal magnitude has higher reproducibility is less affected by noise (no threshold should be selected in order to calculate it).*	11.5	38.5	19.2	-	-	30.8	50
fMRI#26. *It is worth testing, and considering whether or not the final laterality index (LI) for a given individual should be the average of LIs based on voxel count and amplitudes of neural activity.*	7.7	42.3	11.5	15.4	-	23.1	50
fMRI#27. *Laterality indices derived from task-related fMRI should be based on a combination of voxel count and t-values.*	3.8	34.6	50	-	-	11.5	72.6
fMRI#29. *It needs to be defined what laterality is atypical.*	19.2	46.2	11.5	11.5	-	11.5	65.4
fMRI#30. *Non-parametric tests are more appropriate than parametric tests when statistically comparing laterality indices between tasks and groups.*	19.2	26.9	26.9	3.8	-	23.1	46.1
fMRI#31. *Laterality indices are continuous interval variables. Basic operations (addition, subtraction, or multiplication) on laterality indices are not meaningful.*	11.5	46.2	7.7	3.8	-	30.8	57.7

Analysis (thresholding)							
fMRI#32. *Unthresholded t-maps, or preferably z-maps, should be used for calculation of laterality indices (LIs) in regions of interest (ROIs). This should be the case especially when LIs are calculated at several levels, including 50 or lower percentage level, with respect to maximally activated voxel, or the X percentage (e.g., the average of 5% of maximally activated) voxels in the image stat map.*	7.7	46.2	15.4	-	-	30.8	53.9
fMRI#33. *Lateralization indices should not be calculated using only a single threshold of significance.*	23.1	50	11.5	3.8	3.8	7.7	73.1
fMRI#34. *As threshold-independent methods can be used: (a) t-weighting method (yielded to unambiguous and stable lateralization across different weighting functions and congruent with other methods in patients) or (b) Bootstrapping.*	11.5	61.5	15.4	-	3.8	7.7	73
fMRI#35. *There are different methods to compute threshold-free laterality indices. A clear rationale must be provided for why a given method is selected within a given study context.*	42.3	50	3.8	-	-	3.8	92.3
fMRI#36. *Threshold-independent laterality indices are recommended; for instance, using methods that consider the statistical distribution of all voxels in each region of interest.*	23.1	46.2	15.4	3.8	-	11.5	69.3
Reporting							
fMRI#37. *Motivate and report the choice of the contrast to assess functional laterality of a given cognitive function.*	53.8	42.3	-	-	-	3.8	96.1
fMRI#38. *The interpretation of laterality indices requires a careful evaluation of the contribution of both hemispheres, since the index reflects an interaction between task and hemisphere.*	42.3	46.2	3.8	-	-	7.7	88.5
fMRI#39. *We should reach consensus on how to refer to 'mixed/bilateral' language dominance.*	23.1	57.7	7.7	11.5	-	-	80.8
fMRI#40. *If you agreed with the previous statement, which of the following options do you prefer?*							
- Bilateral dominance	30.4						
- I do not favour consensus on this statement	13						
- Mixed dominance	39.1						
- Other	17.4						
fMRI#42. *We should not compare laterality indices between tasks with very different baselines.*	11.5	34.6	15.4	15.4	3.8	19.2	46.1
fMRI#45. *To generate categorical indices, a threshold of 0.2 does not always mean the same hemispheric dominance (or bias) across different LI computation methods.*	7.7	53.8	11.5	3.8	-	23.1	61.5
fMRI#46. *In clinical care and in research, the definition of 'mixed' (or bilateral), 'left' and 'right' language dominance should be defined clearly and simply.*	30.8	50	11.5	-	-	7.7	80.8

fMRI#47. *In clinical care, the determination of language dominance should be made with reference to, but not only on the basis of, a laterality index.*	26.9	42.3	11.5	-	-	19.2	69.2
fMRI#48*. In clinical care, when clinical and laterality index-based determinations of language dominance differ (in category or magnitude) the reasons for this should be clearly and simply detailed.*	23.1	57.7	7.7	-	-	11.5	80.8

Supplementary Table 9: Statement ratings for structural MRI. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (🛇). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 20 experts (1 vote = 5.0%).

Statement ratings: Structural MRI	☺☺	☺	😐	☹	☹☹	🛇	☺☺+☺
Spatial normalization and brain (a)symmetry							
SMRI#1. A strong paradigm in neuroimaging is spatial normalization to a brain template. Performing asymmetry studies often imply the use of a symmetric template. All voxel-based measures of lateralization should only be performed after spatial normalization to a symmetric template	15.0	65.0	10.0	5.0	-	5.0	80.0
SMRI#2. Laterality indices based on structural imaging are influenced by the inherent laterality of the template used for registration of individual brains into a standard space. Thus, it is important that size measurements be made, if possible, in the subject's native space.	30.0	40.0	15.0	-	-	15.0	70.0
SMRI#3. Voxel based morphometry techniques basically ignore individual differences in landmark patterns. One should go back to identify the laterality of particular anatomical metrics manually, design new tools for measuring these individual landmarks, or to use probability atlases for those areas known to be highly individual.	5.0	60.0	10.0	15.0	-	10.0	65.0
SMRI#4. Creating sample based symmetrical templates more strongly respects individual differences of the sample compared to using a symmetrical MNI template.	15.0	50.0	10.0	10.0	-	15.0	65.0
SMRI#5. Regional asymmetry measures can be based on asymmetrical atlas definitions when the goal is measuring individual differences but quantifying the population average asymmetry usually requires an artificially symmetrized atlas, or other procedure such as hemispheric co-registration.	5.0	70.0	10.0	10.0	-	5.0	75.0
Comprehensive measurement							
SMRI#7. The measurement of brain structural asymmetries must take into account the dissociation between hemispheric differences in sulci position and tissue compartment density or volume hemispheric difference.	10.0	65.0	10.0	-	-	15.0	75.0
SMRI#8. To evaluate structural asymmetries between the two cerebral hemispheres, it is possible to focus on anatomically defined regions of interest (ROI: gyrus, sulcus, white matter bundle, central grey nuclei, etc.) and to perform volumetric (size, volume, depth, length, count of fibers, etc.) or morphological (folding, shape, etc.) measurements. It is then essential to ensure that these ROI have been defined in a coherent and corresponding manner between the two hemispheres, with a delimitation method that takes into account the overall morphological asymmetries observed between hemispheres (related to petalias and torque, extent and angularity of the Sylvian fissure, cortical thickness, cortical surface, etc.).	25.0	60.0	5.0	-	-	10.0	85.0

SMRI#9. We should reach a consensus on which dependent variables to use to calculate laterality indices, e.g., FA, MD, and others. Maybe it would make sense to always report a set of variables and not cherry pick the one the reaches significance for laterality effects like sometimes seems to be the case in some papers.	15.0	55.0	10.0	5.0	5.0	10.0	70.0
SMRI#11. Surface-based brain structural asymmetries should be privileged over voxel-based morphometry (VBM) ones because they allow the dissociation between cortical surface area and cortical thickness asymmetries pattern and so provides information on structural asymmetries that may be specific to a given anatomical feature, as opposed to VBM asymmetries analysis that captures information about grey matter volume that is, by definition, the product of CSA and CT.	20.0	20.0	20.0	20.0	10.0	10.0	40.0
SMRI#12. To manually or semi-automatically define regions of interest (ROI) on which to perform structural asymmetry measurements (e.g. volumetry and morphometry with anatomical MRI, microstructure characterization with diffusion or relaxometry MRI), it may be relevant to return some right and left hemispheres in a blind and random way between subjects, so that the operator does not know whether it is one hemisphere or the other (which could bias the ROI definition and therefore the measurements performed). In the event that ROI must be defined manually, several experimenters should delineate all right and left regions, for all subjects, to ensure that asymmetry measurements are not operator-dependent.	15.0	55.0	15.0	-	-	15.0	70.0
SMRI#13. For region-based laterality index calculation, functional/structural regions should be defined for each hemisphere separately based on local anatomical/functional properties.	10.0	60.0	10.0	5.0	-	15.0	70.0
SMRI#14. Since inter-hemispheric correspondence between the left and right cerebral hemispheres is established based on the high-dimensional non-rigid surface registration, laterality index can be computed across the whole cerebral surface on a vertex-by-vertex basis. In particular, we should examine the positional brain surface asymmetries measured as displacements between corresponding vertex pairs in relation to an estimated mid-sagittal plane, which provide a picture of the morphological asymmetry of cerebral surface in great detail. The computation of vertex-wise laterality index can be conveniently extended into any other measurement associated with the surface vertex, such as gyrification index.	-	45.0	20.0	-	5.0	30.0	45.0
SMRI#15. For measurements of volumetric or morphological asymmetries, anatomical MRI techniques (T1 or T2 weighted images) can be used to delineate regions such as gyri or sulci, and diffusion MRI can identify white matter tracts in both hemispheres. In order to assess microstructure asymmetries in both grey and white matter, it is necessary to consider quantitative parameters (e.g., diffusion tensor imaging parameters, relaxation times T1 and T2) that can be reliably compared between the two hemispheres. Indeed, the measurements used to evaluate microstructure asymmetries should not be contaminated by the spatially observed bias in MRI images (especially in T1 and T2 weighted images). This bias can indeed lead to significant signal differences between hemispheres, which are irrelevant and can lead to the interpretation of erroneous asymmetries.	10.0	40.0	15.0	5.0	-	30.0	50.0

SMRI#17rw. Statistics with volume/size-based laterality indices need to be adjusted for brain size. A consensus should be reached on how to achieve this.	25.0	35.0	10.0	20.0	-	10.0	60.0
SMRI#18rw. Volumetric region-specific laterality indices are more meaningful when accounting for the overall size difference between the two hemispheres. A consensus should be reached on how to achieve this.	5.0	60.0	15.0	5.0	-	15.0	65.0
SMRI#19. We need a common approach to report and analyse within-person change in asymmetry over time.	5.0	45.0	30.0	-	-	20.0	50.0

Supplementary Table 10: Statement ratings for fTCD. Percentage of experts rating each statement on a 5-point scale: strongly agree (☺☺), agree (☺), neutral (😐), disagree (☹), strongly disagree (☹☹) and no opinion (🛇). Summed strongly agree and agree votes indicate general level of statement consensus. Results are based on votes from 13 experts (1 vote = 7.7%).

Statement ratings: Functional transcranial Doppler ultrasonography	☺☺	☺	😐	☹	☹☹	🛇	☺☺+☺
Study set-up							
fTCD#1rw. *The standard error of the LI will depend on the number of trials/epochs (N), and is proportional to the square root of N. For fewer than 16 trials, the LI is likely to be unreliable. This also implies exclusion of participants with fewer than 16 trials per condition.*	7.7	61.5	15.4	15.4	-	-	69.2
fTCD#2. *Different types of trial should be presented randomly where possible (e.g., when the instructions are the same and the participant is blind to the manipulation).*	15.4	61.5	-	23.1	-	-	76.9
fTCD#5. *The participant's behaviour during the POI being analysed should be assessed when possible. For this reason, we suggest that overt/active tasks should be favoured when possible.*	7.7	84.6	-	-	7.7	-	92.3
fTCD#6rw. *In-doppler' compliance measures, such as a later report period, add a memory component to the task, invite verbal rehearsal strategies, and do not necessarily reflect performance in the active period. Comparison between LI during period-of-interest and compliance behaviour in time periods outside of the period-of-interest should be avoided.*	23.1	38.5	23.1	7.7	7.7	-	61.6
fTCD#7rw. *Motor activity (i.e. motor responses) should be assessed during fTCD tasks so that asymmetrical motor activity, which is a potential confound, can be reported*.	23.1	46.2	23.1	7.7	-	-	69.3
Data exclusion							
fTCD#10. *Predetermine and report criteria for excluding or terminating a recording - for example, depth of MCA markedly different from usual, participant not able to remain silent in baseline period, etc.*	38	62	-	-	-	-	100
fTCD#11. *Where a laterality index is based on averaging over several trials/epochs, need clear objective criteria for removal of trials that are outliers.*	69	31	-	-	-	-	100
Data processing							
fTCD#12. *A recommended analysis pipeline including downsampling, normalization, heart cycle integration, epoching, data screening, artifact rejection, and baseline correction should be formulated.*	69	31	-	-	-	-	100
fTCD#13rw. Normalization (standardization) and baseline correction should be performed on a trial by trial basis to *avoid slow changes in blood flow velocity in one or both channels contaminating results.*	38	38	23	-	-	-	76

fTCD#14. *When computing a laterality index, need to report whether the analysis was based on the mean difference between left and right channels, or the peak difference, and to justify this choice.*	85	15	-	-	-	-	100
Timings							
fTCD#18rw. *Trial timings should be standardized for known/widely-used tasks. An optimal duration should be piloted for newly developed tasks (for example, by piloting different durations in the same participants performing the same tasks, and analysing the effect on strength of lateralization, internal consistency, etc.). Report the period of interest.*	38.5	46.2	15.4	-	-	-	84.7
fTCD#19. *For each task used with fTCD we should decide on a preferred time-frame around the duration of the baseline and normalization periods.*	23.1	61.5	15.4	-	-	-	84.7
fTCD#20. *Researchers should develop standard methods for measuring LI with fTCD, and should agree on the appropriate POI. If using the same task as a prior study, then the same POI should be used, unless clear justification is given for using different values.*	62	38	-	-	-	-	100
Reporting							
fTCD#22. *Describe max gain used, power settings of doppler... (this can be supplementary material).*	53.8	23.1	15.4	-	-	7.7	76.9
fTCD#23rw. *Gender, age, handedness, history of neurological disorders, degree of bilingualism/multilingualism, medications, of the participants should be reported, as well as information on how they were recruited. Please list additional variables you deem important here (or you think should be removed) to the comments section of this statement.*	30.8	61.5	7.7	-	-	-	92.3
fTCD#24. *Participants' cognitive abilities relevant to the task used should be reported.*	23.1	30.8	30.8	7.7	-	7.7	53.9
fTCD#26. *Report the number of trials/epochs presented per condition and report the cut-off for excluding participants for too few trials.*	85	15	-	-	-	-	100
fTCD#27. *When reporting laterality indices based on task related changes in cerebral blood flow velocity (CBFV), we should report both direction and degree of the hemispheric perfusion difference (the index of lateralization).*	85	15	-	-	-	-	100
fTCD#28. *Group based analyses of Laterality Indices should be accompanied by the relevant Grand Average of the fTCD cerebral blood flow velocity change relative to baseline in the middle cerebral artery (MCA).*	23.1	61.5	7.7	-	-	7.7	84.6
fTCD#29*. We should always report the confidence interval of the hemispheric perfusion difference (the index of lateralization).*	69.2	23.1	7.7	-	-	-	92.3
fTCD#30. *A scatterplot showing individual LIs per task should be included in the manuscript or supplementary materials. These should include an indication of variability such as error bars for standard deviation or standard error across trials.*	53.8	38.5	7.7	-	-	-	92.3

fTCD#31. *Stripped data files (e.g. .exp files with no identifying information) should be routinely supplied with manuscripts, not only when requested*	46.2	30.8	15.4	7.7	-	-	77
fTCD#32. *Raw data must be accompanied by a csv file or similar with analysis parameters such as trial start and end times, trigger channels, number of trials. We could begin to follow the Brain Imaging Data Structure (BIDS) guidelines.*	53.8	46.2	-	-	-	-	100
fTCD#34rw. *To favour comparability, strength of lateralization for newly developed language tasks should ideally be presented together with LIs for a version of a fluency task standardized across the field.*	46.2	53.8	-	-	-	-	100