Annex to:

EFSA AHAW Panel (EFSA Panel on Animal Health and Welfare), More S, Bicout D, Bøtner A, Butterworth A, Calistri P, Depner K, Edwards S, Garin-Bastuji B, Good M, Gortàzar Schmidt C, Miranda MA, Nielsen SS, Sihvonen L, Spoolder H, Willeberg P, Raj M, Thulke H-H, Velarde A, Vyssotski A, Winckler C, Cortiñas Abrahantes J, Garcia A, Muñoz Guajardo I, Zancanaro G and Michel V, 2017. Scientific Opinion on the low atmospheric pressure system for stunning broiler chickens. EFSA Journal 2017;15(12):5056, 85 pp. https://doi.org/10.2903/j.efsa.2017.5056

Annex A – Welfare hazards associated with alternative stunning method in broiler chicken - expert judgement and statistical analysis

Report (Final Draft 25.8.2017)

NP/EFSA/ALPHA/2017/12

Hans-Hermann Thulke, Martin Lange; PG Ecological Epidemiology (EcoEpi) at the Helmholtz Centre for Environmental Research GmbH – UFZ, Dept. Ecological Modelling

1. Objective

The purpose of the study was the technical support regarding a problem on welfare aspects of LAPS in broiler chicken: To evaluate to what extent the LAPS system proposed for stunning poultry is able to provide a level of animal welfare at least equivalent to that ensured by the currently allowed methods for stunning broiler chicken. EFSA provided a list of hazards to broiler chicken welfare in relation to stunning by electrical waterbath (W), acid gas mixtures (G) and LAPS (L). The objective was to provide a quantitative evaluation of the hypothesis that hazards to broiler chicken welfare associated with L are worse/nor worse than those hazards associated with stunning of broiler chickens with the already legislated methods summarized as W and G.

2. Approach

Data driven comparison of the stunning methods was perceived by EFSA as limited because only for the 'L' method sufficient quantitative data were provided. For the alternatives 'W' and 'G' only few data aspects had comparable quality. Therefore, it was agreed to use expert knowledge external to EFSA in order to evaluate the list of hazards to broiler chicken welfare at stunning. The list of hazards identified by EFSA was subjected to individual expert consideration for a ranking by the expected consequence given the animal would be exposed to a certain hazard of the list. In line with Animal Welfare science (AW), the ranking should be performed taking the magnitude of the consequences in to account. Therefore, the experts involved in the study would need both the frequency and duration of exposure to the hazard during standard industrial stunning of broiler chicken. The experts were expected to integrate these data into the magnitude of the consequences resulting from these exposure scenarios and rank the hazards accordingly. As the outcome an ordinal scale was sought for the hazards ranking these according to the associated welfare consequence. Outcome was expected to facilitate the statistical testing of the hypothesis that LAPS hazards by tendency rank worse than the hazards involved the other stunning methods. Rejection of the hypothesis would imply that LAPS is at least not worse to the already practiced methods regarding the involved hazards to the stunned animals' welfare.

3. Methods

a. Data

The study uses the following list of hazards as identified by EFSA:

Annex A - Table 1: List of identified hazards associated with alternative stunning methods in industrial stunning of broiler chicken assuming full compliance with technical protocol i.e. no fault. (L) LAPS, (W) Electrical water-bath, (G) gas mixtures incl. CO2.

METHOD	HAZARD
L	Gas expansion in body cavities / internal organs
L	Removal of air
L	Decreasing air humidity
L	Noise
W	Unintended electric shock
W	Neck cutting
W	Bleeding
W	Handling
W	Hanging and compression of the legs
G	Acidic gas or gas mixture
G	Respiratory stimulant gas or gas mixture
G	Tipping / Tilting

Secondary data are recorded during the expert judgement procedure. These are the rank position assigned to either hazard item in the final ordered list of each individual expert plus certain verbal comment.

The data recording involved three technical parts: (i) a web-based elicitation platform allowing the physical ranking of the identified hazards blinded by stunning methods, (ii) expert invitation und conduct, (iii) data analysis including data aggregation by stunning method, estimation of summary ranks and testing of differences in ranking tendency between stunning methods i.e. median rank using Wilcoxon rank sum test.

b. Roodle-Webinterface

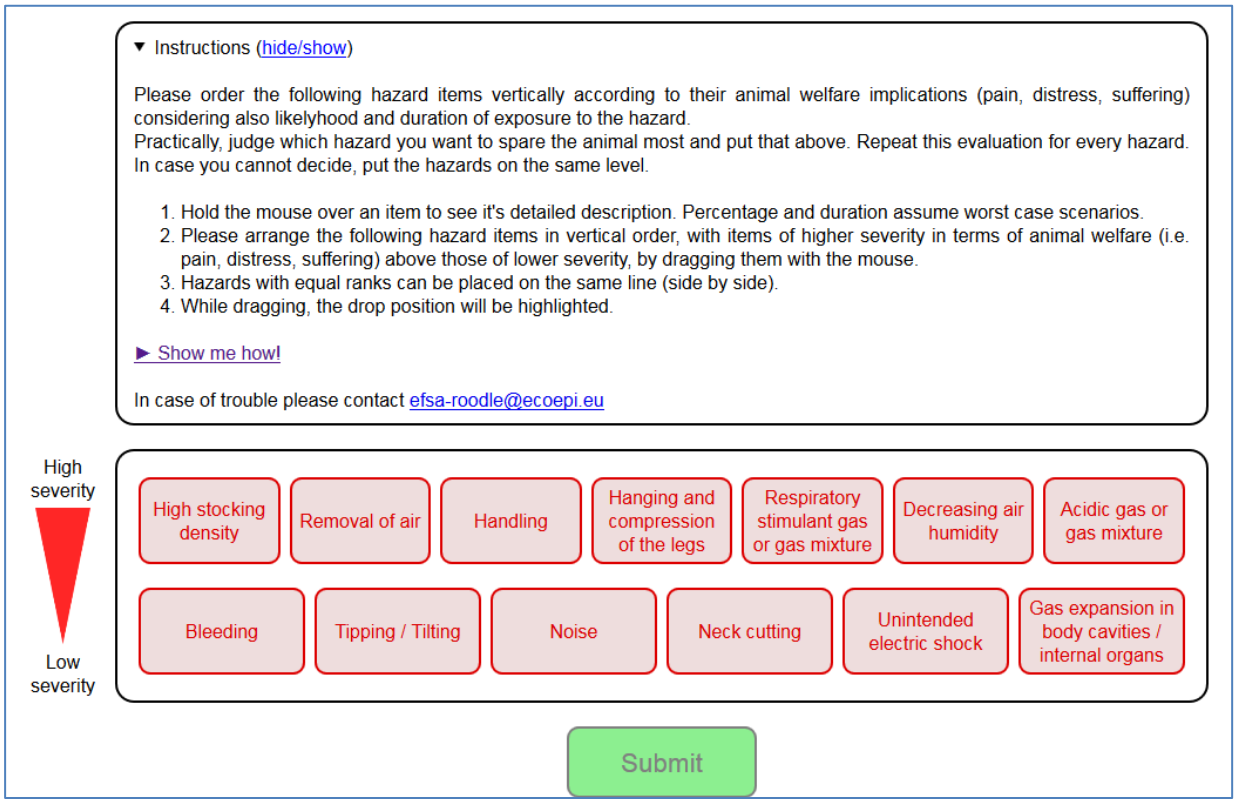
The intention of the exercise was to elicit a ranking of the hazards by expert judgement. The method is established in social science to elicit perceptions of people about certain list of concepts describing different facets (here: hazards to broiler chicken welfare) of a problem (here: welfare at stunning for slaughter). Ranking was sought by the differences in severity of welfare consequences (pain, distress, and suffering) due to the given individual hazards. The duration of exposure and the population percentage exposed were prescribed by EFSA and shared with the judges. The ranking methodology is particular purposeful if the differences between the items to rank is hard to quantify while the superiority of certain items over others is still adequately imagined from expert knowledge. The experts are expected to integrate thus providing the conceptual interface linking causes and consequences.

The ranking method to elicit preference or maximum relevance ordinal scales from complex conceptualisations is developed for the physical move of written items e.g. item-wise paper

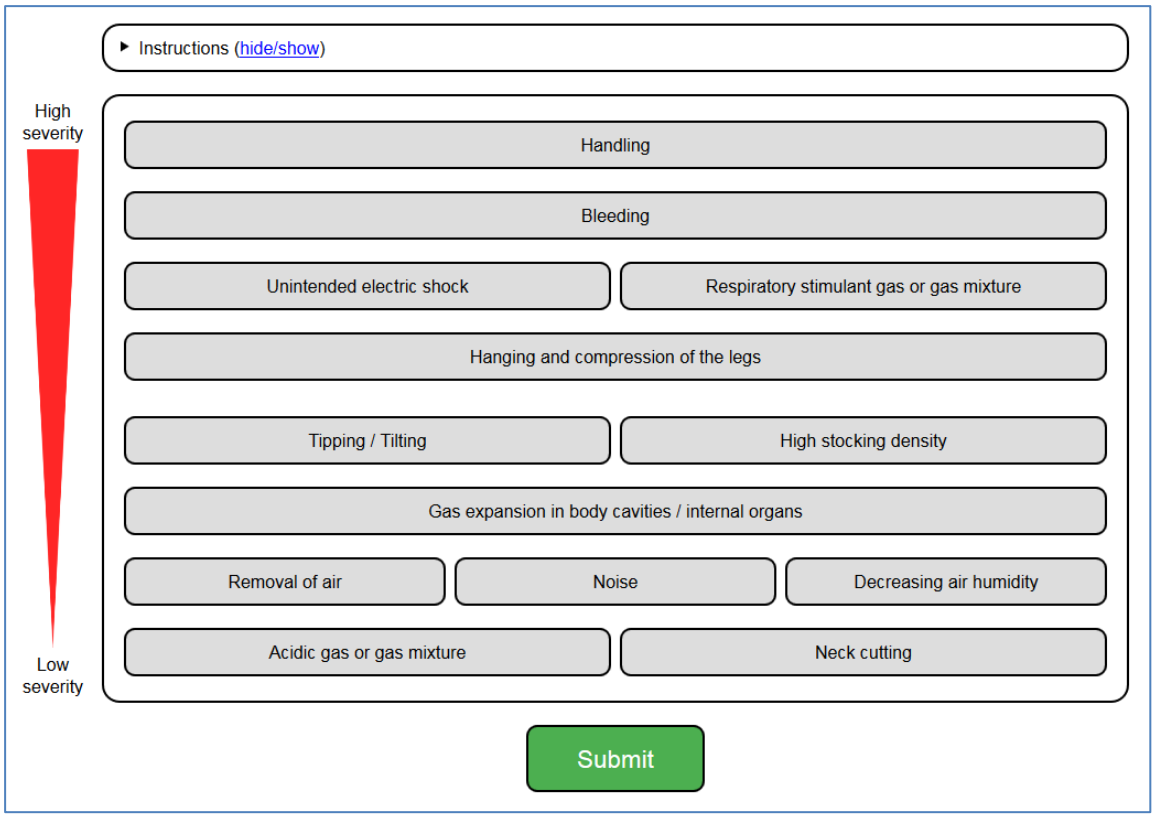
prints. However, the physical meeting with all experts was not logistically possible and therefore online elicitation remained the only option.

To facilitate the adequate experimental setting a web application (**r**anking-d**oodle,** referred to as 'roodle' hereinafter) was developed and implemented in 'php' using SQL database. The roodle application provides the virtual opportunity to the individual user to drag single items, to read background and accompanying data, and to move and drop single items up or down until its position fits the user perception (see Appendix I – Figure 1 and 2). The user can change the ranking ad libitum. Equal level ranking is facilitated in roodle. Only if all items were selected at least once the user was enabled to submit the ranking as final. Multiple submissions are excluded via personalised links which expire after first submission of a final ranking.

The roodle application on time provides the list of personalised links used or pending, the tabulated individual ranking orders submitted and an ad-hoc statistic of average rank per item.



Annex A - Figure 1: Initial screen of the roodle application prior to any action.



Annex A - Figure 2: Arbitrary final arrangement of items. Each item is addressed once (i.e. all greyed). Submission bottom enabled.

During the process a mouse-over displays the input data per hazard including name of hazard, definition, frequency and duration of possible exposure. After the submission the user receives a conformation and the possibility to provide a free text based comment to the evaluator.

c. Expert invitation

All experts received a standardised invitation letter explaining the request and the approach. Further each expert received a personalised link to join the judgement (see Annex A – Appendix (iii)). Experts were accessed again 5 days later to encourage participation.

d. Statistics

The ordinal scale of the hazards retrospectively was reassigned with the associated stunning method (W, G, L). Hence rank positions relative to the stunning method could be subjected to hypothesis testing. By chance the order of Ws, Gs and Ls (respective the associated hazards) should be random i.e. providing the same median rank per method cohort. The rank test identifies systematic deviations in the median ranks between the cohorts. The hypothesis tested assumes that LAPS hazards more often rank worth than the hazards of other stunning methods. Rejection of the hypothesis would imply that LAPS is at least equal to the already practiced methods.

Ranks associated with the items result in an ordinal scale. However, the distance between the ranked items is not measurable. Thus, the differences in severity between two items ranked nearby each other will likely differ between pairs of hazards considered. In other words, one cannot assume that the difference between ranks is equidistant even though the numbers assigned are. This is in contrast to interval data, in which the difference between responses can be calculated and the numbers do refer to a measured "something" (Salivan & Artino, 2013).

Per hazard item the individual expert ratings were summed over all experts. The hazard-wise rank sums were associated with the group identification L, W, and G. The respective vectors of

stratified rank sums were used as input in the non-parametric Wilcoxon rank sum test (R function Wilcoxon.test). Nonparametric tests do not make an assumption about the "shape" of the distribution from which the study data have been drawn. Nonparametric tests are less powerful than parametric tests and usually require a larger sample size (n value) to have the same power as parametric tests to find a difference between groups when a difference actually exists (Salivan & Artino, 2013). However, the non-parametric tests can allow tendency testing even if data are non-normal and values have intrinsically no metric meaning.

4. Results

The study involved 12 hazards (Table 1). 32 experts were invited. The elicitation resulted in 19 valid sets of response data out of 19 experts that submitted their evaluation. Additionally, the EFSA WG on LAPS did perform the ranking but these data were not included in the assessment to exclude biased view by knowing the procedures and the survey construction (n=4 out of 5).

Annex A - Table 2: Summary of the survey data (see Annex A Table 1 for full details). The hazards are ordered by the value of the rank sum taken over all 19 responses per hazard item. The column rank shows the rank of the rank sum values. The column Min and Max reflect the same outcome after step-wise exclusion of one individual expert and underpins the robustness of the general outcome.

Method	Hazard	Rank estimate (n=19)	Min	Max
W	Unintended electric shock	12	11	12
W	Neck cutting	11	11	12
W	Bleeding	10	10	10
G	Acidic gas or gas mixture	9	8	9
W	Hanging and compression of the legs	8	8	9
G	Respiratory stimulant gas or gas mixture	7	6	7
L	Gas expansion in body cavities / internal organs	6	6	7
L	Removal of air	5	4	5
W	Handling	4	4	5
G	Tipping / Tilting	3	3	3
L	Noise	2	1	2
L	Decreasing air humidity	1	1	2

Table 2 summarizes the descriptive outcome of the expert rankings. The final order is according to the total rank sum over all experts per hazard item. The rank order of these sums are given as integer representing the final rank estimate in the third column (i.e. avoiding over interpretation of meaningless differences in the rank sum values resulting from an ordinal scale).

The resulting ranking is robust across the survey experts. Single deletion of experts did change the order of the direct neighbours: Unintended electric shock vs. Neck cutting; Acidic gas or gas mixture vs. Hanging and compression of the legs; Respiratory stimulant gas or gas mixture vs. Gas expansion in body cavities; Removal of air vs. Handling; and Noise vs. Decreasing air humidity.

The ranking of hazards according to the rank sum over all experts is shown in Table 2. The outcome indicates the observation of rankHaz(L) < rankHaz(G) < rankHaz(W). Median hazard are W=10; G=7; L=3.5 (Table 2).

Further testing was using the formal rank-sum test by Wilcoxon against the hypothesis that L hazards are ranked worse by the experts than the W and G hazards.

Tested are the number of all pairs (x[i], y[j]) for which y[j] is not greater than x[i]. The null hypothesis is that the probability of observing a randomly selected value from the first group (i.e. x values) that is larger than a randomly selected value from the second group (i.e. y values) equals one half (i.e. by chance). Testing may be understood as checking for median difference between the two set of value x and y.

We apply the estimated rank sum over all the participants for the alternatives "laps" (Gas expansion in body cavities/internal organs; Removal of air; Decreasing air humidity, Noise), "noLaps" ("water" + "gas") with "water" (Unintended electric shock; Neck cutting; Bleeding; Hanging and compression of the legs; Handling) and "gas" (Acidic gas or gas mixture; Respiratory stimulant gas or gas mixture; Tipping/Tilting).

Wilcoxon rank sum test

data: laps and exist: W = 28, p-value = 0.025

difference in location 5; 95% CI: (1, Inf)

The main contribution is by the hazards related to W. Limiting the test to L>W still rejects the hypothesis of Hazards-L ranked > Hazards-W

data: laps and water: W = 18, p-value = 0.032

difference in location 6; 95% CI: (2, Inf)

Any other test is not providing statistical support for the apparent observation of ordered median ranked hazards W>G(>)L

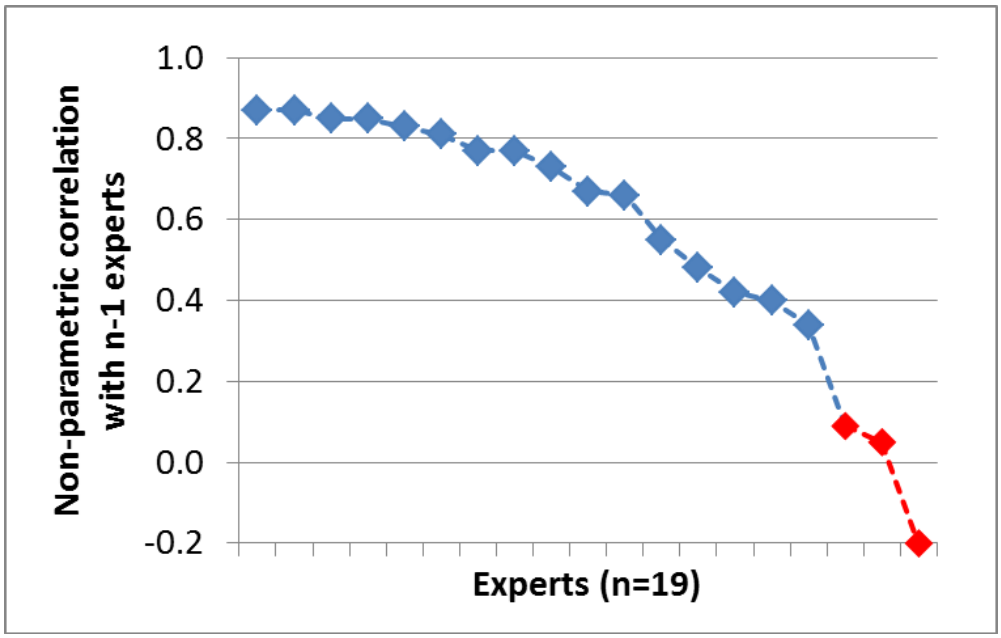
data: laps and gas: W = 10, p-value = 0.115

data: gas and water: W = 12, p-value = 0.125

The hypothesis of homogeneous ranking of the hazards between Laps and Water-bath; or Laps and any existing method was rejected, confirming the subordinate ranking of Laps hazards as statistically significant at 95% probability. The limitation to differentiate between laps and gas is due to the small sample size of 3 vs. 4 hazards.

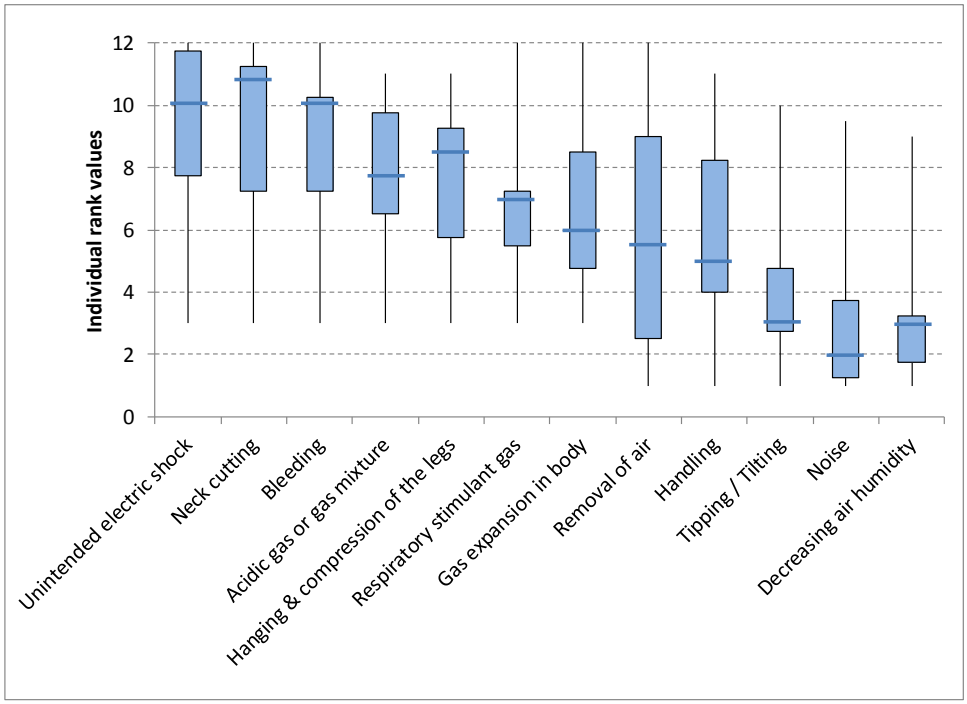
5. Secondary analyses

Noteworthy the ranking of particular hazards was rather homogeneous throughout the majority of experts (Annex A - Fig. 3). There were three experts that did show minor level of agreement on the level of individual rankings of hazard items (expert G, J, K in Table A1 of the Appendix).



Annex A - Figure 3: The concordance (non-parametric correlation) of each individual expert with the remaining set of n-1 experts. Red values indicate non-conformity.

The individual rankings did show a relevant spread although the final outcome is maintained (Annex A - Fig. 4).



Annex A - Figure 4: Distribution of individual ranks (n=19) per hazard item (k=12). The hazards are ordered from left to right reflecting the summary ranking according to Table 2. The box is constructed of the median (central bold line); 25% and 75% quartiles (the box) as well as minimum and maximum the (whiskers)

6. Conclusions

- Expert ranking of hazards associated with three alternative stunning methods (L, W, G) resulted in median ranks per group indicating L<G<W i.e. hazards associated with water-bath stunning had greater median rank compared to gas stunning and LAPS.

- The difference is significant for the hypothesis test against L<(W+G) and L<W.
- However, the relation L<G between the median ranking of hazards associated with L and G cannot be demonstrated as statistical support on 95% level due to the limited number of hazards. This is a methodological issue.
- According to the expert ranking of hazards, the animal welfare outcome is considered to be better under the LAPS method when compared to the electrical water-bath stunning.
- The expert ranking of hazards indicated an animal welfare outcome under the LAPS method not worse compared to gas stunning methods (excluding inert gases). Statistically this relation could not be demonstrated, due to lack of statistical power. This is most likely a methodological issue.
- The LAPS method can be considered to be at least equivalent to one of the currently available stunning methods.

7. References

Sullivan GM, Artino AR. Analyzing and Interpreting Data From Likert-Type Scales. Journal of Graduate Medical Education. 2013;5(4):541-542. doi:10.4300/JGME-5-4-18.

Annex A - Appendix (i)

Appendix (i) Table 1: Set of the response data (n=19). Expert H commented not to use the ranking of "Acidic gas or gas mixture" and therefore the original value was replaced by the mean rank (7.8).

Hazard	CLASS	A, B, C, D, E, F, G, H, I, J, K, L, M, O, P, Q, R, S, T
Unintended electric shock	W	12; 11; 12; 11.5; 7.5; 10; 7; 9.5; 12; 12; 3; 11; 9; 5; 9.5; 8; 11.5; 12; 6
Neck cutting	W	10.5; 7; 11; 11.5; 7.5; 12; 3; 11.5; 11; 3.5; 3; 11; 11; 6; 9.5; 11; 11.5; 8.5; 11.5
Bleeding	W	10.5; 7; 10; 10; 5.5; 7.5; 5; 11.5; 10; 3.5; 3; 11; 10; 8; 9.5; 12; 10; 8.5; 11.5
Hanging and compression of the legs	W	5; 3; 9; 5; 9; 9; 7; 9.5; 8; 11; 10.5; 8.5; 6; 7; 9.5; 5.5; 9; 11; 4.5
Handling	W	3; 7; 4; 4; 4; 11; 9.5; 3.5; 5.5; 9; 5; 8.5; 4; 1; 9.5; 5.5; 8; 1.5; 4.5
Acidic gas or gas mixture	G	8; 11; 7; 7.5; 10.5; 3; 11; **7.8**; 9; 6.5; 7; 6.5; 5; 10; 3.5; 9.5; 6; 10; 9
Respiratory stimulant gas or gas mixture	G	7; 7; 5; 7.5; 5.5; 3; 7; 6; 7; 6.5; 10.5; 6.5; 12; 11; 3.5; 9.5; 4; 5.5; 7
Tipping / Tilting	G	4; 3; 3; 3; 2.5; 7.5; 9.5; 1; 5.5; 10; 1; 1.5; 3; 3; 3.5; 2; 7; 4; 3
Gas expansion in body cavities / internal organs	L	6; 7; 8; 6; 10.5; 3; 12; 6; 3; 5; 10.5; 4.5; 8; 12; 3.5; 4; 5; 7; 9
Removal of air	L	9; 11; 1; 2; 12; 3; 1; 8; 3; 1.5; 10.5; 4.5; 7; 9; 3.5; 7; 2; 5.5; 9
Noise	L	2; 3; 2; 1; 1; 6; 2; 3.5; 1; 8; 7; 1.5; 2; 4; 9.5; 1; 3; 1.5; 1
Decreasing air humidity	L	1; 1; 6; 9; 2.5; 3; 4; 2; 3; 1.5; 7; 3; 1; 2; 3.5; 3; 1; 3; 2

Appendix (i) Table 2: List of raw comments provided with the individual judgements

I could not really place the exposure to acid gas because it all depends on the actual mixture, ie concentration of acid gas and the presence of O2 or other gases. So please ignore the item. [Edit note: This was implemented in the analysis]
Some hazards are difficult to interpret. CO2 at high % are higher in ranking as suggested here. A short electric shock can induce a long effect.
There are many different qualities of the Hazards described, e.g. Tilting (with falling, which I thought of when ranking) is not like tilting and slipping down on slides. This would be ranked below noise. I do not feel very well being forced to reduce to such a ranking and not being able to know more differentiated about the conditions, e.g. concentrations
It was difficult to class these hazards but I think the worst are those relating to loss of breathing. Thank you for indicating the duration of these hazards, it helps to relativize the gravity.
An hazard that was not included for ranking is when poultry is not effectively stunned or "skips" the waterbath stunner and enters the scalding tank conscious. When pigs enter the scalding tank alive the effect is quite disturbing for anyone present (even more for the pig of course) so this should only happen very rarely because of corrective action. With poultry non-effective stunning can easily leave them conscious but immobilised. Was this hazard something that was considered for inclusion and left out? Or was it forgotten?
[Off topic] In my opinion EFSA should include in this list ritual slaughter without stunning in poultry.
[Off topic] Maybe it would be easier to understand the HACCP based welfare assessment system if we consider the hazards and their source, for example, we could have a restricted list of hazards (pain, distress, fear) and at each process step describe the source of the hazards (distress and pain due to rude hanging, fear due to excessive noise during unloading...). Then list the control measures based on the source of the hazard, do the analysis (is it a CCP or not), then put in place if necessary monitoring and corrective actions. With such a method, it would be a HACCP system perfectly similar to the food one and easier for welfare inspectors, and food hygiene managers who implement animal welfare policies in slaughterhouses.

Annex A - Appendix (ii)

Response to secondary request on expert bias due to professional background

According to the evaluation of the study it was asked to test (a) the effect of background (i.e. Academia vs Veterinary Official) on the response rate; and (b) on the ranking outcome.

Ad (a): The requestor assigned the background to each expert asked for participation (i.e. 32). Taking these data together with the response received led to the following cross-table

	Response Yes	**Response No**	
Academia	**12**	**8**	20
Veterinary Official	**7**	**5**	12
	19	13	32

Chi-Squared test on A vs V in terms of response rate did not contradict the hypothesis that experts from academia and official veterinarians were equally willing to respond. The chi-square statistic is 0.009. The p-value is .926. This result is not significant at $p < .05$.

Ad (b): Comparing ranking in A vs V		**SUM**		**RANK**	
		A	**V**	**A**	**V**
Unintended electric shock	W	103.5	76	12	12
Neck cutting	W	103	68.5	11	11
Bleeding	W	99.5	64.5	10	10
Acidic gas or gas mixture	G	93.28	54.5	9	8
Hanging and compression of the legs	W	90	57	7	9
Respiratory stimulant gas or gas mixture	G	87	44	6	7
Gas expansion in body cavities / internal organs	L	90.5	39.5	8	5
Removal of air	L	82	27.5	5	3
Handling	W	67	41	4	6
Tipping / Tilting	G	43	34	**3**	4
Noise	L	43	17	**3**	1
Decreasing air humidity	L	36	22.5	1	2

Median rank values (L; G; W): in A = 4; 6; 10 vs V = 2.5; 7; 10

Both suggesting the tendency of H(L)<H(G)<H(W)

The cohort based evaluation of ranking statistics provided the following output:

Academia	**Official veterinarians**
Wilcoxon rank sum test **data: laps and exist: W = 25.5, p-value = 0.063** difference in location 3.7; 95% CI: (-1, Inf) **data: laps and water: W = 17, p-value = 0.056** difference in location 4.5; 95% CI: (-1, Inf) data: laps and gas: W = 8.5, p-value = 0.240 data: gas and water: W = 12, p-value = 0.125	Wilcoxon rank sum test **data: laps and exist: W = 31, p-value = 0.005** difference in location 6; 95% CI: (3, Inf) **data: laps and water: W = 20, p-value = 0.008** difference in location 7; 95% CI: (4, Inf) data: laps and gas: W = 11, p-value = 0.058 data: gas and water: W = 13, p-value = 0.072

On the statistical certainty level of 90% both cohorts confirm the results found for the complete sample. At 95% the lower ranking of LAPS hazards would not lead to significance in the cohort of experts from academia.

Annex A - Appendix (iii)

Letter to the experts

Dear Expert,

The European Food Safety Agency (EFSA) is asking for your support regarding the evaluation of a set of hazards, for the welfare of broiler chickens, identified by selected experts in the context of different slaughtering procedures. The available scientific literature does not provide comprehensive and harmonised data on the behavioural and physiological responses of animals exposed to the hazards, at this time. Therefore, in absence of sufficient scientific evidences, EFSA needs to elicit the required information from your specialised expertise acquired on these topics.

In this exercise you are asked to assess the hazards in terms of welfare consequences (i.e. pain, distress and suffering) and to order them depending on their severity, given the chance that an animal will be exposed to the hazard and given the duration of this exposure. It is important to clarify that this exercise is NOT about ranking stunning methods or slaughtering processes, but rather about ordering individual hazards. The outcome of this exercise run by you will be a list of prespecified hazards ordered by severity of their welfare consequences after comparison of each hazard against each other.

http://ecoepi.eu/efsa/?id=example.ex@domain.dom

You are provided with a personalised link (see here above and at the bottom of this message) to participate in the assessment which will guide you to an online working desk. Here you will find all the hazards identified by the EFSA's experts. Each hazard is accompanied by a detailed description and quantitative information about likely exposure and duration (worst case scenario). The task is – in short – taking each hazard by dragging it with your mouse and drop it in a new position according to your personal judgement on its consequences on the welfare of the animal. Finally, you may achieve a vertical list of hazards with those with most severe welfare outcomes placed above (towards the head of the list) and those implying less severe outcome placed below (towards the foot of the list). You will be able to move the items up and down until you are satisfied with the ordering. After you submit your results (by clicking the 'Submit' button) you can place comments or annotations for EFSA and the evaluation team.

In case you submit your results by mistake, be aware that the assessment can only be repeated upon request of reactivation to the Administrator at efsa-roodle@ecoepi.eu (submission implies the inactivation of your personalised link).

The judgement will realistically take about 10-15 min and there will be no questions or tasks other than the two described above (I.e. ordering and comments, if any).

While realising that this is short notice EFSA would be extremely grateful if you could provide your input by 21 July the latest. We apologise for this short notice but due to the short time provided to EFSA for the finalisation of his Scientific Opinion we would like to ask you to comply with this timeline if at all possible but definitely without an excessive delay.

Your contribution is crucial to improve the scientific outcome of this exercise and we trust that the short time needed to complete this ranking exercise will contribute to your willingness to respect the timeline and possibly to provide your input before the deadline.

EFSA would like to thank you in advance for your invaluable contribution to this study . FSA would like to reassure you about the anonymity of your specific contribution which will only contribute to the set of evidences for the subsequent steps of the EFSA assessment.

With my very best regards,

EFSA