Supplementary material 4: the statistical tests used to verify the equality between the semiskilled and crowd datasets to the expert's dataset

We interrogate the mean and mode severity level values of replies to each question in the two surveyed samples and examine the correlation of their means and variances with the adjacent parameters of the experts' opinion dataset. To test these correlations, we use the statistical paired t-test to verify the mean differences between the examined datasets and a two-tailed F-test to check the equality of the two adjacent variances. The variance test is an important addition to the mean testing by serving as a proxy for the distribution difference between the two datasets. The null hypothesis for a two-tailed t-test is that the mean severity level difference is equal to zero, while the alternative hypothesis is that the mean severity level difference is not equal to zero, as follows:

$\mu_{diff} = \mu_1 - \mu_2$

H_0: $\mu_{diff} = 0$

H_a: $\mu_{diff} \neq 0$

Whereas μ_1 denotes the mean range of the values evaluated by the experts, μ_2 denotes the mean range of a second tested dataset (whether it is the graduate student sample or the online survey dataset). The difference between these two is μ_{diff}.

The two-tailed F-test is used to assess whether the variances of two examined datasets reflect equality of the populations. Accordingly, we define the following null and alternative hypothesis:

$\sigma^2_{diff} = \sigma^2_1 - \sigma^2_2$

H_0: $\sigma^2_{diff} = 0$

H_a: $\sigma^2_{diff} \neq 0$

Whereas σ^2_1 is the variance of the experts' dataset, σ^2_2 corresponds with the other tested dataset, and σ^2_{diff} is the difference between the two variances (Peck et al., 2008).

Prior to performing the t- and F-tests, one should look for inconsistencies in the examined datasets, as well as verify the normality distribution assumption. The descriptive statistics of the datasets are presented in Table S2. The smallest IQR (1.75) belongs to the mode severity levels of both the

graduate student and online crowd datasets, although the experts IQR is quite close (1.88). The IQR of the other two samples is above 2. The standard deviation (Sd = ~1.8 and below) and the standard error (Stderr = ~0.55 and bellow) of all five distributions is not large in relation to their mean values, thus implying a normal distribution. This is also the indication when comparing the median values of the five distributions to the corresponding mean values (all are equal to or less than ~0.5 levels). The Shapiro-Wilk normality tests presented in Table 3 correspond to the difference between the experts' sample and a given examined sample (mean or mode values). The results of all tests are significant (p-values > 0.05), indicating that one cannot reject the inherent assumption of normality—that is, the yielded differences are all normally distributed (Shapiro & Wilk, 1965). Finally, the assumption of normality is also supported by the QQ-plots presented in Figure 3, in which the four differences accord with the diagonal theoretical normal line within a 95% confidence level.

		Graduate students				**Online crowd survey**			
Q Id	**Experts**	**Mean**	**Mean.Diff**	**Mode**	**Mode.Diff**	**Mean**	**Mean.Diff**	**Mode**	**Mode.Diff**
q1(p)	8.0	6.3	**1.7**	7	**1**	6.8	**1.2**	7	**1**
q2	6.0	5.8	0.2	6	0	6.8	-0.8	7	**-1**
q3	9.5	9.0	0.5	8	1.5	9.3	0.2	10	-0.5
q4	8.0	8.4	-0.4	7	**1**	9.7	**-1.7**	9	**-1**
q5	6.5	6.3	0.2	6	0.5				
q6	9.0	7.9	**1.1**	6	**3**				
q7(p)	8.5	7.7	0.8	8	0.5	8.4	0.1	8	0.5
q8	5.0	4.9	0.1	7	**-2**				
q9	7.5	7.4	0.1	7	0.5	7.4	0.1	7	0.5
q10	7.0	6.6	0.4	7	0				
q11	8.0	7.0	**1.0**	7	**1**				
q12(p)	10.0	10	0.0	10	0	10.6	-0.6	12	**-2**
q13	5.5	5.5	0.0	6	-0.5				
q14	6.0	5.9	0.1	6	0	6.2	-0.2	6	0
q15	9.0	8.9	0.1	11	**-2**				
q16	7.5	6.9	0.6	7	0.5	7.2	0.3	7	0.5
q17	7.5	8.3	-0.8	10	**-2.5**				
q18(p)	8.5	8.8	-0.3	9	-0.5	9.0	-0.5	8	0.5
N	**18**					**610**			
Mean	7.61	7.37		7.5		8.17		8.1	
Sd	1.40	1.45		1.54		1.45		1.79	
Stderr	0.330	0.342		0.364		0.459		0.567	
LCL	6.92	6.65		6.73		7.13		6.82	
UCL	8.31	8.09		8.27		9.21		9.38	

Median	7.75	7.2		7		7.95		7.5	
Min	5	4.9		6		6.2		6	
Max	10	10		11		10.6		12	
IQR (Q1–Q3)	1.88	2.4		1.75		2.28		1.75	

Table S1: Mean and frequent severity level of damage estimation of each surveyed question (graduate students and online crowd survey) in comparison with the evaluation of the experts. Columns: **Q Id** = question id; **Experts** = evaluation made by the authors of this paper (M.Z. and A.S.); **Mean** = mean severity level evaluation of the surveyees; **Mean.Diff** = difference between experts and surveyees evaluation; **Mode** = frequent (mode) severity level of surveyees; **Mode.Diff** = difference between experts and frequent surveyees evaluation.

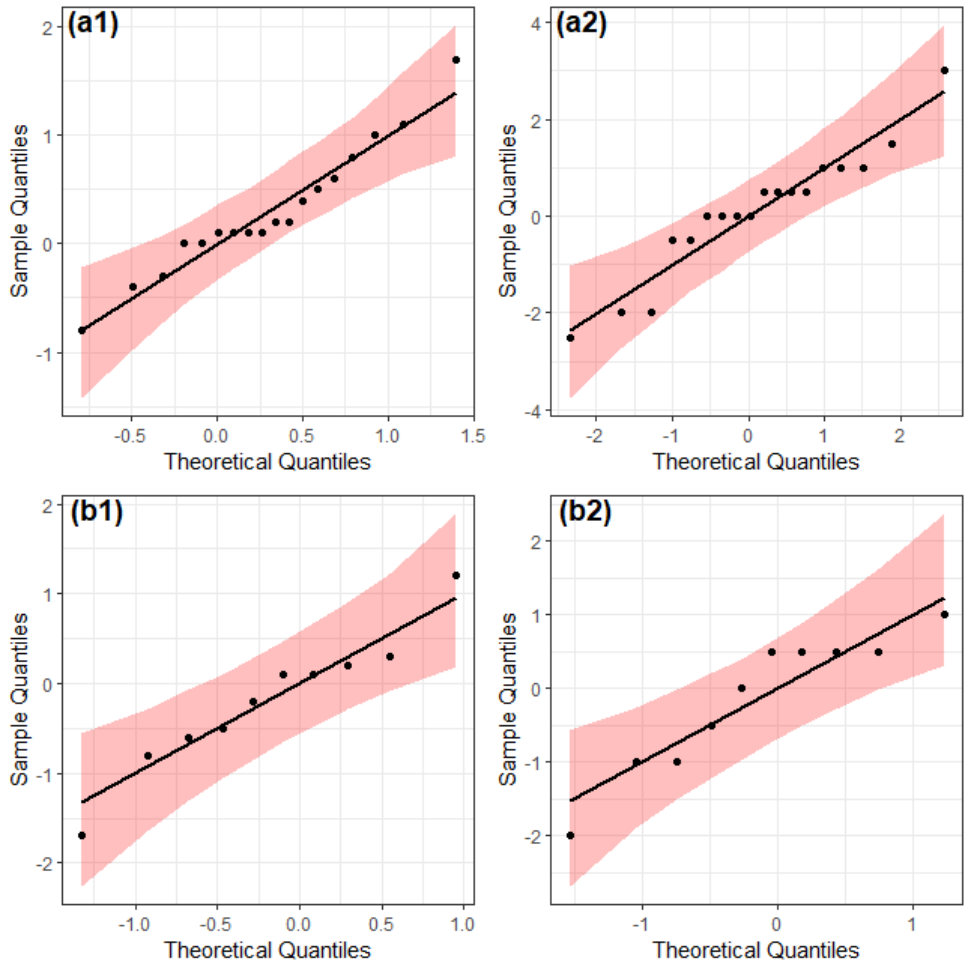

Figure 1: QQ-plot representing the differences between the mean and mode values of the two surveyed datasets with the adjacent values of the experts' dataset: (a1) experts vs. the means of graduate class; (a2) experts vs. the modes of graduate class; (b1) experts vs. the means of the online survey; and (b2) experts vs. the modes of the online survey.

The above verifications support the normality distribution assumption of the four comparisons between the evaluations of the experts and the mean and mode values of the graduate student and online survey datasets. Upon normality verification, one can perform the t- and F-tests as well as inspecting the correlation ratio between the datasets. Table 3 presents the statistics and results of the implemented comparisons (Ids 1–4). The t-test p-value of the four comparisons, excluding the expert-graduate student mean difference, is above 0.05 and, likewise, the F-test p-value of all is above 0.05 (both within a confidence interval of 0.95). That is, excluding the expert-graduate student mean comparison, one cannot reject the null hypotheses (i.e., there is no difference between the mean and variance of the examined datasets), indicating that the given examined datasets are not statistically different from one other. The highest Pearson correlation value is achieved for the expert-graduate student mean (0.91), while the correlations of the expert-graduate student mode, experts-online mean, and experts-online mode are 0.6, 0.85, and 0.86, respectively. The mean difference between the expert-graduate student mode (0.11) and expert-online mode (-0.15) is the smallest, but these values are deduction results of negative and positive differences—that is, an over- and underestimation of the severity level of damage in comparison to the experts' evaluation. The absolute mean difference presents the opposite trend, whereas the expert-graduate student and expert-online means have the smallest absolute mean difference.

Id	Comparison	Shapiro-Wilk test	Shapiro-Wilk p-value	Pearson	t-test	DF	p-value	Confidence interval	Lower	Upper	Mean	Mean.Abs	F-test	p-value
1	Experts-graduate mean	0.958	0.571	0.91	2.16	17	0.045	0.95	0.007	0.592	0.3	0.46	0.988	0.98
2	Experts-online mean	0.952	0.691	0.85	-0.91	9	0.383	0.95	-0.763	0.323	-0.22	0.54	0.803	0.75
3	Experts-graduate mode	0.931	0.209	0.6	0.356	17	0.726	0.95	-0.547	0.769	0.11	0.94	0.821	0.69
4	Experts-online mode	0.894	0.189	0.86	-0.50	9	0.627	0.95	-0.825	0.552	-0.15	0.75	0.6	0.36
5	Experts-graduate mean (q6–q15 substitution)	0.929	0.189	0.92	1.86	17	0.079	0.95	-0.031	0.520	0.244	0.41	0.929	0.882

Table S2: Comparison between experts' evaluations and mean and mode evaluations of the graduate students and online crowd surveys. The Shapiro-Wilk test indicates that the five differences between the datasets are normally distributed while the Pearson correlation is high for all comparisons (≥ 0.6).

REFERENCES

Peck, R., Olsen, C., & Devore, J. L. (2008). Introduction to statistics and data analysis (3rd ed., international student ed. ed.). Thomson Brooks/Cole.

Shapiro, S., & Wilk, M. (1965). An analysis of variance test for normality. Biometrika, 52(3), 591-611.