Do experts agree when assessing risks? An empirical study

Nektarios Karanikas, Steffen Kaspers
Amsterdam University of Applied Sciences / Aviation Academy
Weesperzijde 190
1097 DZ, Amsterdam, The Netherlands

Abstract

Risk matrices have been widely used in the industry under the notion that risk is a product of likelihood by severity of the hazard or safety case under consideration. When reliable raw data are not available to feed mathematical models, experts are asked to state their estimations. This paper presents two studies conducted in a large European airline and partially regarded the weighting of 14 experienced pilots' judgment though software, and the calculation of agreement amongst 10 accident investigators when asked to assess the worst outcome, most credible outcome and risk level for 12 real events. According to the results, only 4 out of the 14 pilots could be reliably used as experts, and low to moderate agreement amongst the accident investigators was observed. Although quite alarming results, this paper does not aim at raising concerns about the skills of experienced employees; rather, we urge organizations to comprehend the distinction between experience and expertise, and focus on training their stuff in published expert judgment methods.

Keywords: expert judgment, risk assessment, risk matrix

1. Introduction

1.1 Background

Every company deals with a variety of risks regardless the field of its operations. Whatever the hazards (e.g., flaws internal to the system, environmental factors) the idea is that if risk is not controlled, it will lead to minor or major losses such as injuries and fatalities, damage in infrastructure and equipment, and decreased customer satisfaction and market share. Safety risk management refers collectively to a process through which organizations aim at eliminating or mitigating hazards, thus reducing their exposure to risks.

The typical risk management cycle consists of hazard identification, risk level assessment, prioritization and implementation of risk controls, monitoring of residual and new risks, and evaluation of preventive measures' effectiveness. The use of risk matrices has been established across many industry sectors though standards and best practice [e.g., 1, 2, 3, 4]. Those matrices are based on the concept that risk is a product of likelihood by severity of each hazard; within the matrix, hazards and threats are placed in a specific cell which corresponds to a particular risk level. The matrix cells are divided into coloured areas that depict the magnitude of risk. Based on the risk level and area, a decision is made about the acceptance, rejection or control of the risk with the introduction of a variety of barriers and defences (e.g., procedures, training, technology).

The use of risk matrices is accompanied by both advantages and disadvantages. The illustration through risk cells and areas has been negatively criticized because it depicts risks in one-dimension [5]. Although a risk matrix is easy to use due to its graphical and seemingly easy layout, sometimes risk matrices offer low resolution, which may result in difficulties when trying to place a risk in the right segment [5, 6]. Smith, Siefert, & Drain [7] argued that viewing the consequence as a single point in a matrix might be problematic since the same situation might happen again but with implications of different magnitude. Duijm [5] concluded that a matrix might be used differently across professionals, some of them considering the most likely scenario and others thinking about the worst case; the aforementioned author concluded that the manner of representation affects how people accept risk. Hubbard & Evans [6] viewed risk matrices as additive or multiplicative scoring methods, which are accompanied by four drawbacks:

- Their use is subject to cognitive biases, as also Smith et al. [7] statistically confirmed.
- The assignment of probability and severity labels is not standardized across the industry and can be changed to accommodate each organization's risk appetite over time.
- The labels assigned to likelihood and severity affect the results themselves (e.g., a 3-point scale provides a different interpretation of risk compared with a 5-point scale).
- There might be correlations which are not visibly taken into account (e.g., cascade failures).

Available raw data from past cases and events is exploited for risk level estimations (e.g., probabilistic calculations, average costs incurred). Support from experts is requested when data about probabilities and outcomes is unavailable, corrupted or unreliable. Nonetheless, the performance of experts in terms of their judgments' accuracy has been questioned; Camerer & Johnson [8] found that simple models outperformed experts, but subsequent research contradicted these findings [9]. So, it is suggested that both, simple models and expert judgment, should be used as complementary to each other in order to merge their advantages [9, 10]. Weighting the experts has been an additional method for collectively eliciting judgements and provide estimations based on the level of expertise offered by each specialist [11].

1.2 Research scope

Taking into account the literature cited above, this paper presents the results of two studies. Part of the objectives of those studies was:

- The assessment of the level of consistency amongst experts when they were asked to assess possible outcomes and risk levels of real events.

- The weighting of experts as means to facilitate decision making when assessing risks.

The studies were performed in a large European airline and the results indicated extremely low agreement amongst the estimations of experts, and their highly uneven weighting.

2. Methodology

As part of their bachelor thesis, Bloemendaal [12] assessed the level of agreement between experts when evaluating risks and Jánossy [13] calculated the weighting of experts when evaluating event probabilities. Both studies were performed at the same large European air operator; the participants of the two studies were different.

2.1 Assessing agreement amongst experts

Bloemendaal [12] presented 12 Air Safety Reports (ASRs) to 10 experienced accident investigators. The company contemplates those employees as experts and asks for their judgment in the frame of safety risk management. The ASRs dated from October 2014 to May 2015 and were stored in the airline's database; ASRs representing event types with the highest frequency were selected. The airline uses a matrix divided into 25 risk levels (5x5 matrix) with 4 risk areas: low – green area, medium – yellow area, high – orange area and substantial – red area (Figure 1). The company had classified the specific ASRs as follows: 3 low, 7 medium and 2 high.

	PROBABILITY				
SEVERITY	A	B	C	D	E
5	🟨	🟨	🟧	🟧	🟥
4	🟩	🟨	🟨	🟧	🟧
3	🟩	🟩	🟨	🟨	🟧
2	🟩	🟩	🟩	🟨	🟨
1	🟩	🟩	🟩	🟩	🟩

Figure 1. The risk matrix type used by the airline.

First, the researcher posed to each expert two open questions for each ASR: "What is the worst outcome?", and "What is the most credible outcome?". Second, the accident investigators assigned to each ASR a risk level in the 5x5 risk matrix, indicating thus the probability and severity level of each event, as well its risk area. Intentionally, the experts were not presented with a predefined list of outcomes, in order to minimize anchoring bias. Their answers were qualitatively analysed in order to develop a mutually exclusive and exhaustively inclusive list of outcomes.

Based on the data collected by Bloemendaal [12], we used the Kendall's W non-parametric test for calculating the inter-rater agreement for the worst, most credible outcome, probability, severity and risk levels, and risk area. Kendall's W ranges between 0 (no agreement) and 1 (complete agreement). The significance level was set at to α=0.05.

2.2 Weighting of expert judgment

Jánossy [13] weighted 14 highly experienced pilots in order to indicate the extent to which the judgment of each expert would be considered when assessing event probabilities. The sample was: 5 pilots flying an A330 aircraft type, 5 pilots flying a B777 aircraft type and 4 pilots flying a B747 aircraft type. The Excalibur software [14] was used for weighting the experts based on seven seed questions; the participants were asked to recall numerical data as follows:

1. IATA flights conducted worldwide two years ago.
2. Hull losses of western-built aircraft occurred per 10 million flights two years ago.
3. ASR submitted the previous year by pilots of the specific airline.
4. ASR of the previous year classified as “High” risk in the airline.
5. ASR of the previous year classified as “Medium” risk in the airline.
6. Take-Off Configuration warnings in the previous year within the airline.
7. Rejected Take-Off at a speed rate higher than 80 knots in the previous year within the airline.

The weights were calculated based on the experts’ performance on the seed questions. Based on suggestions from literature [15, 16] the “Performance Weighting” option of the Excalibur software was preferred [14].

3. Results

3.1 Agreement amongst experts

Tables I and II show correspondingly the list and distribution of the worst and most credible outcome types the accident investigators assigned to the 12 ASRs. The figures in the cells represent the number of experts that attributed the specific outcome to the respective ASR.

Table I: Frequencies of worst outcomes selected per ASR.

	Worst outcome categories										
ASR	Death	Injury, no hospitalisation	Injury with hospitalisation	Hull loss	Loss of control	Runway excursion	Aircraft damage	Mid-air collision	Airprox	Hard Landing	Short landing
1	2	1	5								
2	6	1	1								
3	1			6	2	1					
4			1		2		2	5			
5	1	2		1	4		1				
6	7		2								
7	6	2	1								
8	1			5	1	1			1		
9	1			3	1	1				2	1
10	3						7				
11								10			
12	4		1		4			1			

Table II: Frequencies of most credible outcomes selected per ASR

	Most credible outcome categories											
ASR	Injury, no hospitalisation	Injury with hospitalisation	Hard landing with damage	Loss of control	Damage to aircraft	Hull loss	Mid-air collision	Death	Runway excursion	Long landing	Physical distress	Loss of separation
1	1	1	6									
2	6	2										
3			6	1	2	1						
4	1				3		1					
5					4		1					
6	1	2			2			1				
7		5	3									
8		1	1		3				2	1	1	
9				1		1				5		
10		1			7							
11							5					2
12	1	6		1				1				

Tables III and IV show correspondingly the probability (scale A to E in ascending alphabetical order) and severity (scale 1 to 5 in ascending order), and risk level estimations of experts (i.e. the cross reference of severity and probability levels). Each column of Table IV corresponds to the risk area presented in Figure 1. The numbers in the cells represent how many experts assigned each option (i.e. probability, severity and risk levels) to each ASR. The results for the Krippendorff's Alpha and Freidman tests are presented in Table V.

Table III: Frequencies of probability and severity levels assigned per ASR.

	Probability level					Severity level				
ASR	**A**	**B**	**C**	**D**	**E**	**1**	**2**	**3**	**4**	**5**
1		2		3	5			4	6	
2	2		3	2	2		3	4	1	1
3	1	4	5					5	4	1
4	5	4				3	2	1	2	1
5	4	3	1			2	1	1	4	
6	4	1	3			2	1	1	3	1
7	4	3	2					5	4	
8	6	2	1				1	3	4	1
9	5	3	1	1		1	4	1	3	1
10	5	3	2				1	3	6	
11	7		1							8
12		1	5	2	2			5	5	

Table IV: Frequencies of risk levels assigned per ASR.

	Risk level (for the respective risk area see Figure 1)																								
ASR	**A1**	**A2**	**A3**	**A4**	**A5**	**B1**	**B2**	**B3**	**B4**	**B5**	**C1**	**C2**	**C3**	**C4**	**C5**	**D1**	**D2**	**D3**	**D4**	**D5**	**E1**	**E2**	**E3**	**E4**	**E5**
1									2									1	2				3	2	
2		1		1								1	1	1									3		
3				1				1	2	1			4	1											
4	3	1		1			1	1	1	1															
5	2			2				1	2			1													

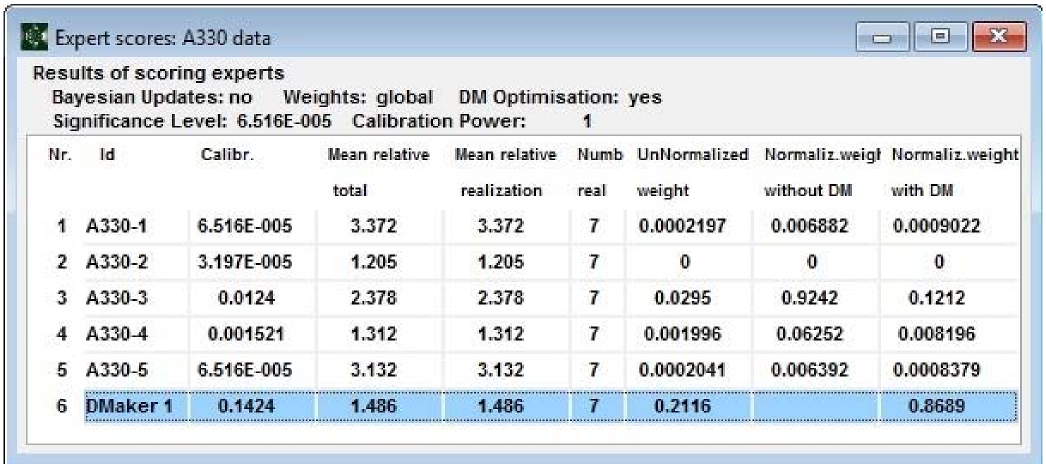
	Risk level (for the respective risk area see Figure 1)																								
ASR	**A1**	**A2**	**A3**	**A4**	**A5**	**B1**	**B2**	**B3**	**B4**	**B5**	**C1**	**C2**	**C3**	**C4**	**C5**	**D1**	**D2**	**D3**	**D4**	**D5**	**E1**	**E2**	**E3**	**E4**	**E5**
6	1	1		2						1	1		1	1											
7			2	2				1	2				2												
8			1	4	1		1	1					1												
9		3	1		1		1		2		1								1						
10		1	1	3				2	1					2											
11					7										1										
12								1					2	3				1	1				1	1	

Table V: Inter-rater agreement results.

Variable	**Kendall's W**	**Significance**
Worst outcome	0.220	0.003
Most credible outcome	0.164	0.027
Probability level	0.305	0.006
Severity level	0.315	0.004
Risk level	0.241	0.000
Risk area	0.550	0.000

3.2 Expert judgment weighting

The results of the Excalibur software are illustrated in Figures 2, 3 & 4 for the A330, B747 and B777 pilots correspondingly. The figures in the column "Normalized Weight without DM" are of interest for the scope of this paper; those represent the proposed weighting of the experts without the advantage of software's Decision Making (DM) function [14].



Expert scores: A330 data

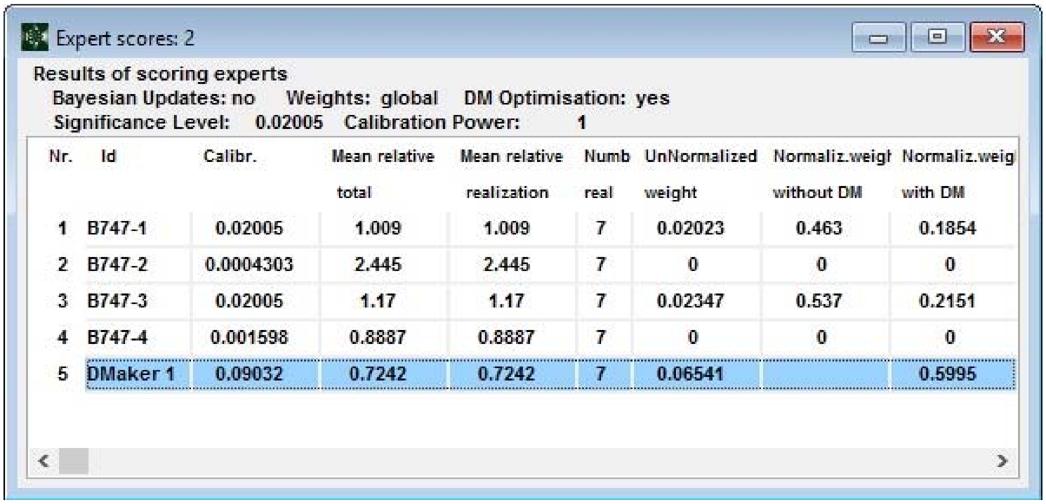
Results of scoring experts

Bayesian Updates: no Weights: global DM Optimisation: yes

Significance Level: 6.516E-005 Calibration Power: 1

Nr.	Id	Calibr.	Mean relative total	Mean relative realization	Numb real	UnNormalized weight	Normaliz.weigl without DM	Normaliz.weight with DM
1	A330-1	6.516E-005	3.372	3.372	7	0.0002197	0.006882	0.0009022
2	A330-2	3.197E-005	1.205	1.205	7	0	0	0
3	A330-3	0.0124	2.378	2.378	7	0.0295	0.9242	0.1212
4	A330-4	0.001521	1.312	1.312	7	0.001996	0.06252	0.008196
5	A330-5	6.516E-005	3.132	3.132	7	0.0002041	0.006392	0.0008379
6	DMaker 1	0.1424	1.486	1.486	7	0.2116		0.8689

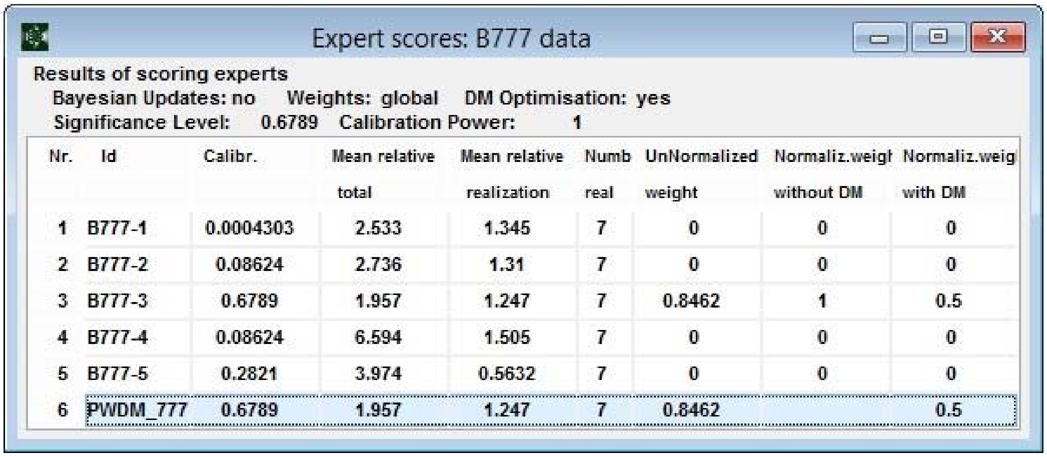
Figure 2. Weighting results for the A330 pilots.



Results of scoring experts
Bayesian Updates: no Weights: global DM Optimisation: yes
Significance Level: 0.02005 Calibration Power: 1

Nr.	Id	Calibr.	Mean relative	Mean relative	Numb	UnNormalized	Normaliz.weigt	Normaliz.weig
			total	realization	real	weight	without DM	with DM
1	B747-1	0.02005	1.009	1.009	7	0.02023	0.463	0.1854
2	B747-2	0.0004303	2.445	2.445	7	0	0	0
3	B747-3	0.02005	1.17	1.17	7	0.02347	0.537	0.2151
4	B747-4	0.001598	0.8887	0.8887	7	0	0	0
5	DMaker 1	0.09032	0.7242	0.7242	7	0.06541		0.5995

Figure 3. Weighting results for the B747 pilots.



Results of scoring experts
Bayesian Updates: no Weights: global DM Optimisation: yes
Significance Level: 0.6789 Calibration Power: 1

Nr.	Id	Calibr.	Mean relative	Mean relative	Numb	UnNormalized	Normaliz.weigt	Normaliz.weig
			total	realization	real	weight	without DM	with DM
1	B777-1	0.0004303	2.533	1.345	7	0	0	0
2	B777-2	0.08624	2.736	1.31	7	0	0	0
3	B777-3	0.6789	1.957	1.247	7	0.8462	1	0.5
4	B777-4	0.08624	6.594	1.505	7	0	0	0
5	B777-5	0.2821	3.974	0.5632	7	0	0	0
6	PWDM_777	0.6789	1.957	1.247	7	0.8462		0.5

Figure 4. Weighting results for the B777 pilots.

According to the results, if the company requested a judgment from those participants the following would apply:

- A330 pilots: The judgment of pilot A330-3 should be taken mostly into account, whereas A330-2 pilot's opinion should not be considered at all.

- B747 pilots: Pilots B747-2 and B747-4 should be excluded and only assessments of pilots B747-1 and B747-3 should be contemplated with about the same weight.

- B777 pilots: Only the opinion of the pilot B777-3 should be counted.

4. Discussion

The results regarding the agreement amongst 10 experienced accident investigators suggested low to moderate agreement in the assessment of all variables considered; Kendall's W ranged from 0.164 for the most credible outcome to 0.550 for the risk area. Even prior to any statistical calculations, the observation of remarkably scattered figures across Tables I, II, III and IV signified a low agreement amongst the experts.

Interestingly, the results of the expert's weighting with the use of the Excalibur software showed remarkable differences amongst the pilots. Out of the 14 participants only 4 would be considered reliable in their judgments, hence decreasing significantly the pool of experts the specific airline could consult if an assessment was necessary. Aggregated results for all 14 pilots were not available in order to examine possible variances amongst all participants. In addition, it must be noted that in all weighting methods the quality of the seed questions plays a paramount role in the results; in the study of Jánossy [13] the effects of the questions used were not exhaustively examined.

5. Conclusions

Certainly, since both studies were conducted in one airline and limitations in the use of the methods employed might exist (i.e. selection of ASRs and effects of seed questions) we do not claim generalizability of the results. Nonetheless, even under those potential imperfections in the quantification of agreement amongst experts, the qualitative evaluation of the data collected confirm the limitations of probability-severity matrices' usage in risk assessments. However, since cognitive biases are inevitably present in each decision and judgment, the goal of this paper is not to raise concerns about the competencies and the trustworthiness of skilful employees.

Organizations need to realise that extensive working experience is not directly associated with expertise [17]. In line with the literature, we propose that companies consider the consistent use of published expert judgment methods and train their safety professionals and experienced staff accordingly; this way, a combination of hard data and human judgment is likely to support effective decision making. In addition, careful interpretation of the results from relevant software and acknowledgement of limitations such software impose will avoid negative implications on the relationships amongst employees and disturbances in organizational culture.

The powerful and unreplaceable human capabilities have been and will always be crucial for maintaining and improving current safety levels. Diversity must be valued when collecting views (e.g. hazard identification, planning of remedial actions against threats). However, when it comes to assigning risk levels in matrices, which prevail the risk decision making across the industry, sufficient consistency and reliability are indisputably required. If the latter cannot be achieved, risk matrices and other probabilistic risk assessment tools must hold only a supportive role in safety risk assessment, and it is rather time to explore the value of alternative methods and tools.

References

[1] AIRMIC, Alarm and IRM, *A structured approach to Enterprise Risk Management (ERM) and the requirements of ISO 31000,* United Kingdom, 2010.

[2] ARMS, *The ARMS Methodology for Operational Risk Assessment in Aviation Organizations,* ARMS Working Group, 2010.

[3] D. Smith, Reliability, Maintainability and Risk: Practical Methods for Engineers, 8th ed., Oxford: Butterworth-Heinemann, 2011.

[4] ICAO, *Safety Management Manual,* Montreal: International Civil Aviation Organization, 2013.

[5] N. J. Duijm, "Recommendations on the use and design of risk matrices," *Safety Science,* vol. 76, pp. 21-31, 2015.

[6] D. Hubbard and D. Evans, "Problems with scoring methods and ordinal scales in risk assessment," *IBM Journal of Research and Development,* vol. 54, no. 3, pp. 2:1-2:10, 2010.

[7] E. D. Smith, W. T. Siefert and D. Drain, "Risk matrix input data biases," *Systems Engineering,* vol. 12, no. 4, pp. 344-360, 2009.

[8] C. F. Camerer and E. J. Johnson, "The process-performance paradox in expert judgment: How can experts know so much and predict so badly?," in *Toward a General Theory of Expertise: Prospects and Limits*, vol. 342, Cambridge, Cambridge University Press, 1991, pp. 195-217.

[9] M. Jørgensen, "Forecasting of software development work effort: Evidence on expert judgement and formal models," *International Journal of Forecasting,* vol. 23, no. 3, pp. 449-462, 2007.

[10] R. T. Hughes, "Expert judgement as an estimating method," *Information and Software Technology,* vol. 38, no. 2, pp. 67-75, 1996.

[11] R. M. Cooke and L. H. J. Goossens, "Expert judgement elicitation for risk assessments of critical infrastructures," *Journal of Risk Research,* vol. 7, no. 6, pp. 643-656, 2004.

[12] M. Bloemendaal, *Increasing objectivity in risk assessments: Integration of accident report data in risk assessment as a means to increase objectivity, Bachelor Thesis (unpublished),* Amsterdam: Amsterdam University of Applied Sciences, 2015.

[13] M. Jánossy, *Expert judgement on accident probabilities in use, Bachelor Thesis (unpublished),* Amsterdam: Amsterdam University of Applied Sciences, 2015.

[14] T. Delft, "Classical Model Software - Excalibur," Lighttwist Software, 2013. [Online]. Available: http://www.expertsinuncertainty.net/Publications/Excalibur/tabid/4386/Default.aspx.

[15] R. M. Cooke, M. E. Wittman, D. M. Lodge, J. D. Rothlisberger, E. S. Rutherford, H. Zhang and D. M. Mason, "Out-of-Sample Validation for Structured Expert Judgment of Asian Carp Establishment in Lake Erie," *Integrated Environmental Assessment and Management vol. 10,* pp. 522-528, 2014.

[16] J. W. Eggstaff, T. A. Mazzuchi and S. Sarkani, "The effect of the number of seed variables on the performance of Cooke's classical model," *Reliability Engineering and System Safety,* pp. 72-82, 2013.

[17] M. W. Wiggins and T. Loveday, Diagnostic Expertise in Orgnizational Enviroments, Surrey: Ashgate, 2015.

[18] K. Krippendorff, "Computing Krippendorff's Alpha-Reliability," 2011. [Online]. Available: http://repository.upenn.edu/asc_papers/43. [Accessed 20 November 2015].

[19] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia Medica,* vol. 22, no. 3, pp. 276-282, 2012.

[20] A. J. Viera and J. M. Garrett, "Understanding Interobserver Agreement: The Kappa Statistic," *Family Medicine,* vol. 37, no. 5, pp. 360-363, 2005.

[21] D. Freelon, "ReCal OIR: Ordinal, interval, and ratio intercoder reliability as a web service," *International Journal of Internet Science,* vol. 8, no. 1, pp. 10-16, 2013.