Evidence assessment schemes for semi-quantitative risk analyses
A response to Roger Flage and Terje Aven

Goerlandt, Floris; Reniers, Genserik

DOI
10.1016/j.ssci.2017.04.008

Publication date
2017

Document Version
Final published version

Published in
Safety Science

Citation (APA)
Goerlandt, F., & Reniers, G. (2017). Evidence assessment schemes for semi-quantitative risk analyses: A response to Roger Flage and Terje Aven. *Safety Science*, *98*, 12-16. https://doi.org/10.1016/j.ssci.2017.04.008

Safety Science 98 (2017) 12–16

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/ssci

Letter to the editor

Evidence assessment schemes for semi-quantitative risk analyses: A response to Roger Flage and Terje Aven

1. Introduction

In a recent letter to the editor, Flage and Aven (2017) provide a number of comments on one of our articles, which focuses on the assessment of uncertainty in risk diagrams (Goerlandt and Reniers, 2016). Their comments mainly address our remarks concerning ambiguity in the qualitative uncertainty assessment scheme proposed in Flage and Aven (2009), but some additional comments are also made on the work by Goerlandt and Montewka (2015a, 2015b), mainly related to the assessment of evidential biases.

We strongly believe that critical reflections on published articles and discussions about fundamental issues are essential to improve the current state-of-art in risk research. One of the writers of the letter to the editor has in earlier work called for such contributions (Aven and Zio, 2014). Considering as well that there currently is no very strongly established culture of discussion on key ideas and concepts in risk research, an issue raised as well e.g. by Rosa (2010), we appreciate the efforts made by the Safety Science Editorial Board to stimulate such discussion by providing a platform for this kind of contributions.

Thus, we are pleased to receive some relevant and thoughtful comments on our work by Roger Flage and Terje Aven. Upon invitation by the Editor, we are happy to continue this discussion by providing a response to their letter.

In the following sections, we will subsequently address three issues raised by Flage and Aven (2017). First, we comment on the partial misrepresentation of the uncertainty assessment scheme for supporting semi-quantitative risk analyses, proposed in Flage and Aven (2009). Second, we address the issue of ambiguity in the original scheme by Flage and Aven (2009). Third, we raise some concerns about the new scheme presented in Flage and Aven (2017). Subsequently, we conclude with some general remarks on evidence assessment.

2. Partial misrepresentation of the uncertainty assessment scheme by Flage and Aven (2009)

Flage and Aven (2017) highlight that in Goerlandt and Montewka (2015a), the uncertainty assessment scheme presented in Flage and Aven (2009) is partially erroneously represented. The original text reads (Flage and Aven, 2009, p.13–14 [emphasis added]):

> The category classifications (minor, moderate, significant) will be case-specific and subject to judgment by the analyst, but the following descriptions could serve as a guideline:
>
> **Significant uncertainty**
> *One or more of the following conditions are met:*
> - The phenomena involved are not well understood; models are non-existent or known/believed to give poor predictions.
> - The assumptions made represent strong simplifications.
> - Data are not available, or are unreliable.
> - There is lack of agreement/consensus among experts.
>
> **Minor uncertainty**
> *All of the following conditions are met:*
> - The phenomena involved are not well understood; the models used are known to give predictions with the required accuracy.
> - The assumptions made are seen as very reasonable.
> - Much reliable data are available.
> - There is broad agreement among experts.
>
> **Moderate uncertainty**
> *Conditions between those characterising significant and minor uncertainty, e.g.:*
> - The phenomena involved are well understood, but the models used are considered simple/crude.
> - Some reliable data are available.

In more recent publications (Aven, 2013; Flage et al., 2014) the evidence assessment focuses on 'strength of knowledge' rather than 'level of uncertainty', where, mutatis mutandis, the same assessment scheme is applied with significant, moderate and minor uncertainty corresponding to weak, moderate and strong strength of knowledge. This conceptual shift away from 'level of uncertainty' of the evidence toward an appreciation of the 'strength of the evidential support' is followed also by Goerlandt and Reniers (2016).

In Goerlandt and Montewka (2015a), the evidence assessment scheme by Flage and Aven (2009) is applied for three purposes: (i) as a basis for selecting risk model elements to prioritize, for incorporating alternative hypotheses within the risk model, (ii) within an assessment scheme for an expert-driven deliberation on the effect of assumptions on the model-based risk quantification, and (iii) as a global evidence assessment related to the overall outcome uncertainties of the events in focus in the risk analysis. The former two uses are part of the first risk analysis stage, which focuses on risk modelling and an expert deliberation and review, whereas the latter is part of the second risk analysis stage, which focuses on deliberative judgment and is oriented towards decision makers. In Goerlandt and Reniers (2016), the evidence assessment scheme is elaborated somewhat, accounting for a set of qualities of the different evidence types, similarly as in Goerlandt and Montewka (2015a, 2015b).

http://dx.doi.org/10.1016/j.ssci.2017.04.008

Flage and Aven (2017) rightfully remark that in the paper by Goerlandt and Montewka (2015b), the scheme presented in Flage and Aven (2009) is erroneously represented in the sense that the criterion for significant uncertainty is not that *all* of the conditions are met, but *one or more* of the conditions. We apologize for this mistake.

In Goerlandt and Reniers (2016, p.70), the uncertainty rating scheme by Flage and Aven (2009) is represented verbatim. However, elsewhere (Goerlandt and Reniers, p. 72 [emphasis added]), we state:

> Another issue [...] is the ambiguity in the definitions of the rating schemes as proposed in Flage and Aven (2009) and Amundrud and Aven (2012). [...] This is due to the fact that the combinations of the phrases "all of the following conditions..." and "one or more of the following conditions..." for *low and high uncertainty* and the phrase "conditions between those characterising low and high..." [...] lead to not mutually exclusive categories.

Here, Flage and Aven (2017) do not comment explicitly on the issue of which exact phrase applies for the criteria for low and high uncertainty, but instead focus on our claim about the ambiguity related to the non-exclusivity of the categorization, which we will address below. Nevertheless, we wish to clarify here that the phrases of the criteria are also not exactly the same in Flage and Aven (2009) and Amundrud and Aven (2012). Table 1 clarifies the issue for the relevant literature covered in Goerlandt and Reniers (2016) and Flage and Aven (2017), showing the phrases used for high and low uncertainty (strong and weak strength of evidence) rating. In all cases, uncertainty/strength-of-knowledge is classified as medium with conditions between the other two ratings.

From the overview, it is seen that the phrasing suggested in Flage and Aven (2009) has been used most, with exceptions in Amundrud and Aven (2012) and Goerlandt and Montewka (2015a). It is also clear that there is a trend towards focusing on strength of evidence rather than on uncertainty related to the evidence as a basic concept. Given these above differences in the literature, we appreciate the letter by Flage and Aven (2017) to consolidate and clarify their current interpretation on the issue.

3. Ambiguity in the original uncertainty assessment scheme (Flage and Aven, 2009)

We believe that the wording applied in the qualitative assessment scheme is important. As evident from the quote from Goerlandt and Reniers (2016), see Section 2, our main concern is that the linguistic ambiguity can lead to a situation where the categories are not mutually exclusive, i.e. that they are not clearly delineated. This is undesirable, as possibly different interpretations can render the evidence assessment scheme unreliable in practical applications, which negatively affects the reliability of the risk analysis if the evidence assessment is used to adjust the rating of the risk event, as e.g. in Amundrud et al. (2013, p. 204):

> From this we that the risk events are first categorized as high, medium or low with respect to probabilities and consequences. Then, the risk events are adjusted one category up if the strength-of-knowledge is classified as medium or weak.

We are sympathetic to the theoretical discussion that risk analyses in most cases (especially those for low-probability/high-consequence events) cannot be considered reliable tools in the sense intended in Aven and Heide (2009), as found in a literature review by Goerlandt et al. (2017). Nevertheless, we believe it to be a worthwhile endeavour to aim to minimize the unreliability, e.g. by avoiding linguistic ambiguity in qualitative schemes. Also, given earlier arguments about the importance of clear interpretations of the tools used to measure uncertainty quantitatively (Aven, 2011), we take it that the same should apply to qualitative ratings of uncertainty and strength of evidence in a risk analysis context. We understand this view is shared in Flage and Aven (2017).

An example makes clear why there is, in our interpretation, ambiguity in the assessment scheme by Flage and Aven (2009). Take the following case where a risk event is characterized based on following evidence:

- The assumptions made are seen as very reasonable.
- Data are not available.
- There is broad agreement/consensus among experts.
- The phenomena involved are not well understood, and the models used in the analysis are believed to give poor predictions.

Interpreting the phrasing in Flage and Aven (2009), see Section 2, it is certain that this case does not meet the requirements for a 'minor uncertainty' rating, because of the unavailability of data and poor predictions obtained by the model. One assessor could interpret the scheme to say that there is 'significant uncertainty' because *one or more* of the conditions for this category are met: (1) there is no data available and (2) the models used are believed to provide poor predictions. Another assessor could meaningfully argue that this is a case of 'moderate uncertainty', because while there is no data available and the models are believed to provide poor predictions, the assumptions made are seen as very reasonable and there is broad agreement among experts. This could be interpreted as a

Table 1
Phrases applied for the categorization of the evidence uncertainty | strength of evidence.

Score		Phrase	Focus on U or SoE?	Source
Uncertainty (U)	Strength of evidence (SoE)			
Low	Strong	All of the conditions are met	U	Flage and Aven (2009)
High	Weak	One or more of the conditions are met		
Low	Strong	One or more of the conditions are met	U	Amundrud and Aven (2012)
High	Weak	One or more of the conditions are met		
Low	Strong	All of the conditions are met	SoE	Aven (2013)
High	Weak	One or more of the conditions are met		
Low	Strong	All of the conditions are met	U	Abrahamsen et al. (2014)
High	Weak	One or more of the conditions are met		
Low	Strong	All of the conditions are met	SoE	Flage et al. (2014)
High	Weak	One or more of the conditions are met		
Low	Strong	All of the conditions are met	U	Goerlandt and Montewka (2015a)
High	Weak	All of the conditions are met		
Low	Strong	All of the conditions are met	SoE	Goerlandt and Reniers (2016)
High	Weak	One or more of the conditions are met		

case where the conditions are between those characterising significant and minor uncertainty. As these are two plausible interpretations of the case based on the phrasing given in the uncertainty assessment scheme by Flage and Aven (2009), this is a case of linguistic ambiguity as understood in Johansen and Rausand (2015).

In Flage and Aven (2017), the ambiguity in the original uncertainty assessment scheme of Flage and Aven (2009) is acknowledged, and we appreciate that our feedback about this has been positively received. We also are sympathetic to the clarified interpretation of the evidence assessment scheme in Flage and Aven (2017), where each of the evidence aspects is classified as either strong, moderate or weak. A similar approach has been suggested in Goerlandt and Reniers (2016), however without an overall classification where the four evidence attributes are subsequently combined as in Flage and Aven (2017).

4. Some concerns about the new evidence assessment scheme by Flage and Aven (2017)

Flage and Aven suggest an updated interpretation of the evidence assessment scheme, which is introduced as follows (Flage and Aven, 2017):

> [Table 2] shows that when allowing for each of the four aspects ('evidential categories') to be classified as either strong, moderate or weak [...] there is no ambiguity. In [Table 2], the strength of knowledge classification for each of the four aspects, phenomenological understanding/models, data, expert statements and assumptions, are allowed to vary across the categories strong (S), moderate (M) and weak (W) knowledge (corresponding to low, medium and high uncertainty, respectively, in Flage and Aven (2009)). The overall classification resulting in each case based on the criteria in Flage and Aven (2009) is seen to be unique. There is no ambiguity problem present.

We agree that with the above described assessment scheme, a unique overall classification rating is obtained. In that sense, it is an improvement over the scheme presented in Flage and Aven (2009). However, we have some concerns about this new scheme as well, which we believe requires careful reflection and consideration. We address following issues in the following sections: (i) possible ambiguities in the new scheme, (ii) possibly undesirable overall classification of the overall evidential strength, and (iii) the application of the scheme in a risk analysis and risk management context.

4.1. Ambiguity in the new evidence assessment scheme by Flage and Aven (2017)

The claim that there is no ambiguity in the scheme of Table 2 is in our view unsupported and probably incorrect. We raise two issues below: (i) the interpretation of the phrasing for each evidential category, (ii) the somewhat unclear role of the 'assumptions' category in relation to the other three.

If the phrasing from Flage and Aven (2009), shown in Section 2, is applied for each evidential category as in Flage and Aven (2017), it is rather likely that different assessors will understand these phrases differently in a given context, and even that these interpretations may depend on the context in which the assessment scheme is used. Findings by Beyth-Marom (1982) show that this is the case for verbal expressions of probability, based on which we find it a plausible hypothesis that there may be large variation in interpretations of the phrases. For instance in the phrase, "much reliable data are available", there can be different interpretations as to what is 'much data', or as to how reliable the data really is. Similar issues can arise e.g. with phrases like 'a well understood phenomenon', 'the assumptions are very reasonable' or 'broad expert agreement' for the other evidence categories. One major issue here is the fact that the concepts addressed in the phrases (reliability, agreement, understanding) are matters of degree, which may not be amenable to very precise measurement. If such different interpretations indeed occur, this would to our understanding imply there is linguistic ambiguity as intended by Johansen and Rausand (2015).

Another possible source of ambiguity, which we believe needs to be clarified further, is the way that assumptions are treated as a separate category in the assessment scheme proposed in Flage and Aven (2017). When reflecting on the phrase "the assumptions made are seen as very reasonable", it is in our view not fully clear what the assumptions exactly refer to in relation to the other three evidential categories. The issue here is that within the other evidence categories, several assumptions are made as well. Models (in particular engineering models) have been described as "[comprising] at least propositions expressing scientific representations and propositions expressing empirical assumptions" (Diekmann and Peterson, 2013, p.211). At least certain types of data are also gathered based on assumptions in the sense that different frames of reference can lead to different findings and contents of the data. This is for instance the case in accident investigations, where different underlying assumptions (accident models) lead to different causes found for the accident, i.e. different data is gathered, see Lundberg et al. (2009). Finally, expert judgments are also based on a particular background knowledge available to the assessor, which typically contain various assumptions as well (O'Hagan et al., 2006). In the application of the evidence assessment scheme, it is not fully clear if these are the kinds of assumptions which are in focus, or if assumptions refer to other background assumptions outside the space of the already considered data, judgments, or models. In the former case, it is not fully clear how to make a combined judgment on whether the assumptions involve strong/medium/weak evidence, if these are not the same for the considered data, judgments, or models. This issue can in our view also lead

Table 2
Strength of knowledge classification scheme according to Flage and Aven (2017).

No.	Phenomena/model	Data	Expert statements	Assumptions	Overall classification
1	S	S	S	S	S
2–17	M	S/M	S/M	S/M	M
	S/M	M	S/M	S/M	M
	S/M	S/M	M	S/M	M
	S/M	S/M	S/M	M	M
18–81	W	S/M/W	S/M/W	S/M/W	W
	S/M/W	W	S/M/W	S/M/W	W
	S/M/W	S/M/W	W	S/M/W	W
	S/M/W	S/M/W	S/M/W	W	W

to different interpretations of the evidence assessment scheme, which would likewise be a case of linguistic ambiguity in the understanding of Johansen and Rausand (2015).

4.2. Possibly undesirable overall classification of evidential strength in the new scheme

The new scheme follows a combinatorial logic that as soon as one evidence category is medium, the overall classification is also medium, unless there is an evidence category rated as weak, in which case the overall classification is weak. While this is internally consistent, we question the logic behind this.

Take for example the cases E1 to E3 presented in Table 3. Here, there is always a weak data evidence category ('no or unreliable data available', see Section 2), whereas the other evidence categories are all either strong (case E1), medium (E2) or weak (E3). In case of E3, there would probably be broad agreement that the overall evidential strength is weak. However, in case E2, it could be quite meaningfully argued that the overall evidential strength is medium rather than weak, because there are more evidential elements rated as medium. In case E1, we find it a quite plausible hypothesis that many analysts would rate the overall classification as strong. Even if in such a case there is little data available to confirm the model results and expert statements (which would often occur in practical settings), having some - even unreliable - data could in our view even further strengthen the support of the other evidential categories, leading to an overall strong evidential rating. In case the limited, unreliable data would show a large discrepancy with the model results and expert statements, this can be taken as an indication to lower the overall evidential rating to medium (not weak, because the data is not very reliable).

Another issue is illustrated in cases E4 and E5 of Table 3, relating to the case where not all evidential aspects are relevant. This issue is addressed in Flage and Aven (2017, [emphasis added]) as follows:

> [...] the case where not evidential aspects ('evidential categories') are *relevant (available)* had already been identified as problematic and rectified before the publication by Goerlandt and Montewka (2015a). For example, Aven (2014) added the phrase 'whenever they are relevant', to account for cases where not all of the four listed aspects are relevant.

In case E4, an analyst has strong data, expert statements and assumptions available, but makes no use of a model. The overall classification is strong, following the logic of Section 2 and the information in the above quote. In a similar case E5, all else being equal, another analyst additionally applies a crude and simple model, which does not give very good predictions. Now, with the model evidence classified as weak, the overall classification becomes weak. For similar reasons as in case E1, we question if this desirable.

Another example is case E6, where there is no data available, but all other evidential categories are rated as strong. From the above quote, it seems the case can be judged with overall rating strong, as the data is simply discarded. However, considering the scheme of Table 2 and the corresponding phrases of Section 2, an ambiguity arises for the data category. On the one hand, the scheme rates the data category as weak as data is not available, which leads to an overall weak evidence rating. On the other hand, if the data category is disregarded as no data is available, as per the above quote, the overall evidence rating would be strong.

4.3. Application of the new scheme in a risk analysis and risk management context

A fundamental issue is also that the evidence assessment scheme is intended to be used as part of a risk analysis. While we certainly are sympathetic to the view that an assessment of uncertainties (or strength of evidence) should be part of a risk analysis, as evident from our earlier work (Goerlandt and Reniers, 2016), we have some concerns about how exactly the evidence assessment is linked to the risk analysis and subsequent risk management.

The main issue is that in various earlier publications, the evidence assessment scheme has been connected to the risk analysis in a way that with low or medium evidential support, the risk rating is increased to a higher category. This approach has been presented e.g. in Abrahamsen et al. (2014), see Section 3. The same idea has been presented in connection with the 'assumption deviation risk' method, a different method to assess the strength of knowledge, see Aven (2013, p.139):

> Next an overall direct judgment is made of the strength of knowledge for the triplet risk assignments (the assumption considered is the same), using again the strong, medium and weak categories. In the case that a weak or medium score is assigned, the risk score based on the triplet assignment can be moved up one category, from medium to high risk, or from low to medium risk.

We are sympathetic to the idea that weak evidence for risk judgements should be communicated to the decision makers, and also that the risk management process should account for this. This has been argued also elsewhere, e.g. Klinke and Renn (2002) and Kristensen et al. (2006), and we take no issue with this in principle. However, our concern in this regards is the way this is practically implemented.

The evidence assessment scheme shown in Flage and Aven (2017) leads to a situation where for the overall classification, there is only one case corresponding to strong evidence, whereas there are 16 cases corresponding to medium and 64 corresponding to weak evidential strength. If the ideas in Amundrud et al. (2013) and Aven (2013) are applied, this implies that in nearly all cases where all evidential categories are considered (80 out of 81), the risk score is moved up one category. This is shown in Table 4, for a case where the risk information is classified in three categories low, medium and high, as is common in risk matrix approaches (Ale et al., 2015; Duijm, 2015).

From the table, it is evident that most risk events would be classified as high (5 out of 9) or medium (3 out of 9), whereas low risk events would be rare (1 out of 9). Given that the strength of evidence would mostly be weak or medium and only in rare cases

Table 3
Example cases of the strength of knowledge classification presented in Flage and Aven (2017).

Case	Phenomena/model	Data	Expert statements	Assumptions	Overall classification
E1	S	W	S	S	W
E2	M	W	M	M	W
E3	W	W	W	W	W
E4	N/A	S	S	S	S
E5	W	S	S	S	W
E6	S	N/A	S	S	S/W

Table 4
Implication of the evidence rating to the risk event rating.

Risk rating	Strength of evidence		
	Weak	Medium	Strong
Low (L)	M	M	L
Medium (M)	H	H	M
High (H)	H	H	H

as strong, it is clear that the complete risk picture is very likely to be dominated by risk events rated as high or medium.

We find this problematic, for instance because it diminishes the resolution of the risk matrix, which has been raised as a limitation of the typical risk matrices in earlier work as well, see e.g. Cox (2008). This limited resolution can raise problems in practical risk management, as it leads to "risk ties", i.e. situations where qualitatively different risks cannot anymore be distinguished because of the way in which the information is aggregated. It is for instance questionable if the situation where a risk is rated medium with weak strength of evidence requires the same concern and treatment as a risk is rated high with a strong strength of evidence. The authors have commented on the appropriateness of this risk ranking approach also elsewhere, see Goerlandt and Reniers (2017).

5. Discussion and conclusions

With our response to the letter to the editor by Flage and Aven (2017), we have aimed to provide a number of critical reflections on the current state of art in the evidence assessment in a risk analysis context. We have focused on the original scheme by Flage and Aven (2009) and the new scheme presented in Flage and Aven (2017). Notwithstanding our comments, we find these schemes important contributions to the risk research discipline, and we in principle agree with the underlying rationale of the ideas by these authors. Our motivation for providing our reflections and feedback is intended to improve the current methodologies, which we believe to be important.

While we have focused on the two evidence assessment schemes in focus in the letter by Flage and Aven (2017), we are certainly open to critical reflection on, debate about and improvements of our suggested approaches, e.g. the qualitative evidence assessment schemes in Goerlandt and Montewka (2015a, 2015b) and Goerlandt and Reniers (2016). It goes beyond the scope of this letter to address this in detail, but in hindsight we acknowledge that also in the schemes presented there, there may be issues which need clarification and improvement. For instance, in Goerlandt and Reniers (2016), the descriptions of the evidential characteristics may lead to different interpretations by different analysts, and the separate category for assumptions as an evidential category has similar problems as discussed in Section 4.1. Furthermore, the approach in Goerlandt and Reniers (2016) makes the evidential support explicit for the different evidence categories, but does not combine these into an overall rating. This avoids challenges related to how exactly to combine the evidence categories as discussed in Section 4.2, and also considers that different weights could be assigned to the importance of different evidential categories. However, this approach may lead to other challenges e.g. related to the interpretability of the presented information, the perceived practical usefulness, or others.

Dr. Floris Goerlandt
Aalto University, Finland
E-mail address: floris.goerlandt@aalto.fi

Prof. Genserik Reniers
University of Antwerp, Belgium
E-mail addresses: genserik.reniers@uantwerpen.be

References

Abrahamsen, E.B., Amundrud, O., Aven, T., Gelyani, A.M., 2014. Safety oriented bubble diagrams vs. risk plots based on prediction intervals and strength-of-knowledge assessments. Which one to use as an alternative to risk matrices? Int. J. Bus. Continuity Risk Manage. 5 (3), 197.

Ale, B., Burnap, P., Slater, D., 2015. On the origin of PCDS – (probability consequence diagrams). Saf. Sci. 72, 229–239.

Amundrud, O., Aven, T., 2012. A practical guide on how to present and visualize the result of risk and vulnerability analysies in a societal safety and security context. In: Presented at the 11th International Probabilistic Safety Assessment and Management Conference and the Annual European Safety and Reliability Conference, Helsinki, Finland.

Amundrud, Ø., Veland, H., Aven, T., 2013. Risk management recommendations – an alternative to the 22 July report. In: Proceedings of the European Safety and Reliability Conference (ESREL2013), Amsterdam, the Netherlands.

Aven, T., 2011. Interpretations of alternative uncertainty representations in a reliability and risk analysis context. Reliab. Eng. Syst. Saf. 96 (3), 353–360.

Aven, T., 2013. Practical implications of the new risk perspectives. Reliab. Eng. Syst. Saf. 115, 136–145.

Aven, T., 2014. Risk, Surprises and Black Swans: Fundamental Ideas and Concepts in Risk Assessment and Risk Management. Routeledge, New York.

Aven, T., Heide, B., 2009. Reliability and validity of risk analysis. Reliab. Eng. Syst. Saf. 94 (11), 1862–1868.

Aven, T., Zio, E., 2014. Foundational issues in risk assessment and risk management. Risk Anal. 34 (7), 1164–1172.

Beyth-Marom, R., 1982. How probable is probable? A numerical translation of verbal probability expressions. J. Forecast. 1 (3), 257–269.

Cox, L.A., 2008. What's wrong with risk matrices? Risk Anal. 28, 497–512.

Diekmann, S., Peterson, M., 2013. The role of non-epistemic values in engineering models. Sci. Eng. Ethics 19 (1), 207–218.

Duijm, N.J., 2015. Recommendations on the use and design of risk matrices. Saf. Sci. 76, 21–31.

Flage, R., Aven, T., 2009. Expressing and communicating uncertainty in relation to quantitative risk analysis. Reliab. Risk Anal. Theory Appl. 2, 9–18.

Flage, R., Aven, T., 2017. Comments to the article by Goerlandt and Reniers titled "On the assessment of uncertainty in risk diagrams". Saf. Sci. 84, 67–77.

Flage, R., Aven, T., Zio, E., Baraldi, P., 2014. Concerns, challenges, and directions of development for the issue of representing uncertainty in risk assessment. Risk Anal. 34 (4), 1196–1207.

Goerlandt, F., Montewka, J., 2015a. A framework for risk analysis of maritime transportation systems: a case study for oil spill from tankers in a ship-ship collision. Saf. Sci. 76, 42–66.

Goerlandt, F., Montewka, J., 2015b. Expressing and communicating uncertainty and bias in relation to quantitative risk analysis. In: Nowakowski, T., Młyńczak, M., Jodejko-Pietruczuk, A., Werbińska-Wojciechowsja, S. (Eds.), Safety and Reliability: Methodology and Applications. Proceedings of the European Safety and Reliability Conference 2014 (ESREL2014), Wroclaw, Poland, 14–18 September 2014. CRC Press, pp. 1691–1699.

Goerlandt, F., Khakzad, N., Reniers, G., 2017. Validity and validation of safety-related quantitative risk analysis: a review. Saf. Sci. http://dx.doi.org/10.1016/j.ssci.2016.08.023.

Goerlandt, F., Reniers, G., 2016. On the assessment of uncertainty in risk diagrams. Saf. Sci. 84, 67–77.

Goerlandt, F., Reniers, G., 2017. An approach for reconciling different perspectives and stakeholder views on risk ranking. J. Cleaner Prod. 149, 1219–1232.

Johansen, I.L., Rausand, M., 2015. Ambiguity in risk assessment. Saf. Sci. 80, 243–251.

Klinke, A., Renn, O., 2002. A new approach to risk evaluation and management: risk-based, precaution-based, and discourse-based strategies. Risk Anal. 22 (6), 1071–1094.

Kristensen, V., Aven, T., Ford, D., 2006. A new perspective on Renn and Klinke's approach to risk evaluation and management. Reliab. Eng. Syst. Saf. 91 (4), 421–432.

Lundberg, J., Rollenhagen, C., Hollnagel, E., 2009. What-You-Look-For-Is-What-You-Find – the consequences of underlying accident models in eight accident investigation manuals. Saf. Sci. 47 (10), 1297–1311.

O'Hagan, A., Buck, C.E., Daneshkhah, A., Eiser, J.R., Garthwaite, P.H., Jenkinson, D.J., Oakley, J.E., Rakow, T., 2006. Uncertainty Judgments: Eliciting Expert Probabilities. Wiley, p. 338.

Rosa, E.A., 2010. The logical status of risk – to burnish or to dull. J. Risk Res. 13 (3), 239–253.