Safety Science 158 (2023) 105963

Contents lists available at ScienceDirect

Safety Science

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

The use of low-level theory to guide the interpretation of road safety evaluation studies

Rune Elvik, Alena Katharina Høye

Institute of Transport Economics, Gaustadalleen 21, 0349 Oslo, Norway

ARTICLE INFO

Keywords:
Evaluation
Validity
Confounding
Road lighting
Meta-analysis
Systematic variation

ABSTRACT

There are two main interpretations of empirical research: methodological and substantive. A methodological interpretation usually rejects a study by arguing that it is based on poor data or methods. A substantive interpretation accepts results as showing real effects. This paper argues that by developing and testing hypotheses about systematic variation in the effects of road safety measures, it may be possible to defend a substantive interpretation of the results of studies that might otherwise be rejected on methodological grounds. Studies evaluating the road safety effects of road lighting are used to illustrate the approach. Ten hypotheses are proposed and tested by means of two meta-analyses. Most of the hypotheses are supported. Thus, although many studies evaluating the road safety effects of road lighting control poorly for potential confounding factors, the systematic pattern of results found in these studies indicates that they mainly show the effects of road lighting, not of confounding factors not controlled for.

1. Background and research problem

The main objective of road safety evaluation studies is to estimate the effects of road safety measures. The best method for evaluating the effects of a measure is a randomised controlled trial, also known as an experiment. However, very few road safety evaluation studies are experimental (Elvik 2021). Observational studies can never provide as convincing evidence of causality as experimental studies. It would nevertheless be wrong to conclude that all observational studies are worthless. While it is difficult to measure study quality, it is possible to assess systematically whether the results of observational studies are likely to mainly reflect the effects of a road safety measure rather than the effects of randomness, bias or confounding.

This paper uses the validity framework developed by Shadish, Cook and Campbell (2002) to illustrate how the theoretical validity of the results of a set of evaluation studies can be assessed. The validity framework is presented in the next section. Studies evaluating the road safety effects of road lighting are used as a case. Hypotheses about systematic variation in the effects of road lighting are developed. These hypotheses are tested by comparing them to the results of two meta-analyses (Elvik 1995, Høye 2021) summarising the findings of a large number of studies that have evaluated the effects on crashes of road lighting. The main question addressed in the paper is:

Can the results of studies that score low for statistical conclusion validity and internal validity be trusted to mainly show the effects of a road safety measure if the studies score high for theoretical validity?

2. The validity framework for interpreting results of research

Shadish, Cook and Campbell (2002) distinguish between four types of validity. These can be regarded as aspects of study quality, which means the extent to which a study, or set of studies, is free from known sources of error and bias. The higher the validity, the more confident we can be that the results of a study approximate the truth. Definitions of the four types of validity are given in Table 1.

2.1. Statistical conclusion validity

It should be noted that the definitions of statistical conclusion validity and theoretical validity given in Table 1 differ from those given by Shadish, Cook and Campbell (2002). Shadish, Cook and Campbell define statistical conclusion validity as the validity of estimates of statistical relationships between variables. The definition given in Table 1 refers to sampling theory, which forms the basis of statistical inference, e.g. in the form of confidence intervals. In estimating a confidence interval, one relies on methods based on sampling theory, which assume that the data have been randomly sampled from a known population.

This is rarely the case for road safety evaluation studies. These are

E-mail addresses: re@toi.no (R. Elvik), alh@toi.no (A. Katharina Høye).

https://doi.org/10.1016/j.ssci.2022.105963
Received 26 April 2022; Received in revised form 2 October 2022; Accepted 6 October 2022
Available online 15 October 2022

Table 1
Definitions of four types of validity.

Type of validity	Definition
Statistical conclusion validity	Studies have high statistical conclusion validity If their results are representative of a known population; are unbiased as far as the presence of bias can be ascertained; and are as precise as sample size and techniques of analysis allow.
Internal validity	Studies have high internal validity if their results reflect a causal relationship between a treatment and its effects.
Construct validity (theoretical validity)	Studies have high theoretical validity if their results support a falsifiable theory about systematic variation in findings between different treatment conditions.
External validity	Studies have high external validity if their results can generalised to other contexts than those in which the studies were performed.

very often based on convenience samples, whose representativeness of any known or theoretical population is unknown. Moreover, samples are usually small and results therefore highly uncertain. In general, the statistical conclusion validity of individual road safety evaluation studies is low. However, when studies are replicated a large number of times, over a long period and in many countries, it becomes possible to assess the stability of results in time and space. If the results of repeated studies are similar, this indicates that although the data were not obtained by random sampling from a known population, the samples are still sufficiently similar to produce similar results. In this sense, successful replication may compensate for the usually low statistical validity of each study. This point is further discussed in Section 2.4.

2.2. Theoretical validity

Shadish, Cook and Campbell use the term construct validity rather than theoretical validity. They define construct validity as the adequacy of operational definitions of theoretical concepts. In Table 1, theoretical validity is defined as the extent to which a study or set of studies support hypotheses about systematic variation in the effects of a treatment. This definition is based on the recognition that the effects of road safety measures usually vary systematically, depending on characteristics of the measure and the context in which it is used (Hauer et al. 2012). Therefore, the possibility of assessing the theoretical validity of road safety evaluation studies depends on whether hypotheses regarding systematic variation in effects can be formulated and tested empirically.

Theoretical validity can rarely be assessed for a single study. A single study will rarely be large enough to address all sources of systematic variation in effects. It may address some of them. In a large sample, one may, for example, determine if the effects of a road safety measure vary according to crash severity. One may perhaps also determine whether effects vary between groups of road users. Most studies are, however, too small to probe for systematic variation in effects. This variation emerges only at the level meta-analysis, when the results of several studies can be combined. Theoretical validity is therefore mainly relevant for a set of studies, less for a single study.

2.3. Internal validity

Internal validity refers to the basis for inferring a causal relationship between a treatment and its impacts. As noted above, only randomised controlled trials can provide a strong basis for inferring causality. In any observational study, this basis is considerably weaker. The most important aspect of internal validity in observational road safety evaluation studies, is how well studies control for potentially confounding factors. This can have a major influence on estimates of effect; see Elvik (1997) for examples. Different study designs embody different degrees of control for confounding. In evaluations of road lighting, the following main types of study design have been used:

1. Before-and-after studies using crashes in daylight as comparison group.
2. Before-and-after studies using unlit road sections as comparison group.
3. Simple case-control studies, i.e. comparisons of lit and unlit roads that may differ in other characteristics.
4. Case-control studies in which cases and controls have been matched or stratified according their values on potentially confounding variables.
5. Multivariate statistical analyses in which road lighting is one of several variables whose statistical relationship to crashes is estimated by means of regression coefficients.

In general, design 2 controls better for confounding than design 1. Design 4 controls better for confounding than design 3. Design 5 may control better for confounding than all of the other four designs, but it runs the risk of producing erroneous estimates of the effect of road lighting due to endogeneity (Elvik 2011). This means that if road lighting is installed on roads with a bad crash record, these roads may continue to be less safe than other roads despite the fact that road lighting may have reduced the number of crashes. In a statistical analysis based on cross-sectional data, this may result in an erroneous positive regression coefficient for road lighting.

Thus, none of the designs are ideal and all are likely to involve some residual confounding. However, by comparing the results of studies employing different designs, one may assess how robust results are with respect to control for confounding. The concept of "robustness with respect to confounding" can be defined as follows: Results of evaluation studies are robust with respect to confounding if studies with different degrees of control for confounding factors obtain similar results. The robustness of results to confounding has been assessed separately for the studies included in 1995 meta-analysis (Elvik 1995) and 2021 meta-analysis (Høye 2021). Fig. 1 presents the results for the 1995 meta-analysis.

Studies have been classified according to crash severity and study design. Within each group, the two boundaries of the 95 % confidence interval located closest to each other have been identified. Please note that these boundaries are not necessarily based on the same study design. For fatal crashes, the lower 95 % boundary (0.263) was based on before-after studies using crashes in daylight as control. The upper 95 % boundary (0.484) was based on all study designs. It is seen that most of the best estimates of effect lie within the narrowest possible confidence interval, except for estimates referring to crashes of unspecified severity.

Fig. 2 reports an assessment of robustness to confounding in the 2021 meta-analysis.

It is seen that the narrowest confidence intervals are wider than in the 1995 meta-analysis. This reflects the fact that different study designs are associated with a greater variation in estimates of effect than in the 1995 meta-analysis. Nevertheless, a fairly high robustness with respect to confounding remains. It is therefore concluded that although many studies have low internal validity, their findings do not appear to be greatly influenced by confounding factors the studies have not controlled for.

2.4. External validity

As far as external validity is concerned, Elvik (2012) assessed it for studies evaluating the effects of road lighting and found that it was high. This means that the results of evaluation studies have been replicated consistently over time and across countries. In other words: The effects estimated have remained stable over time and hardly varied from one country to another. Thus, as far as external validity can be assessed, it appears to be high. Like theoretical validity, external validity applies to a set of studies rather than a single study.

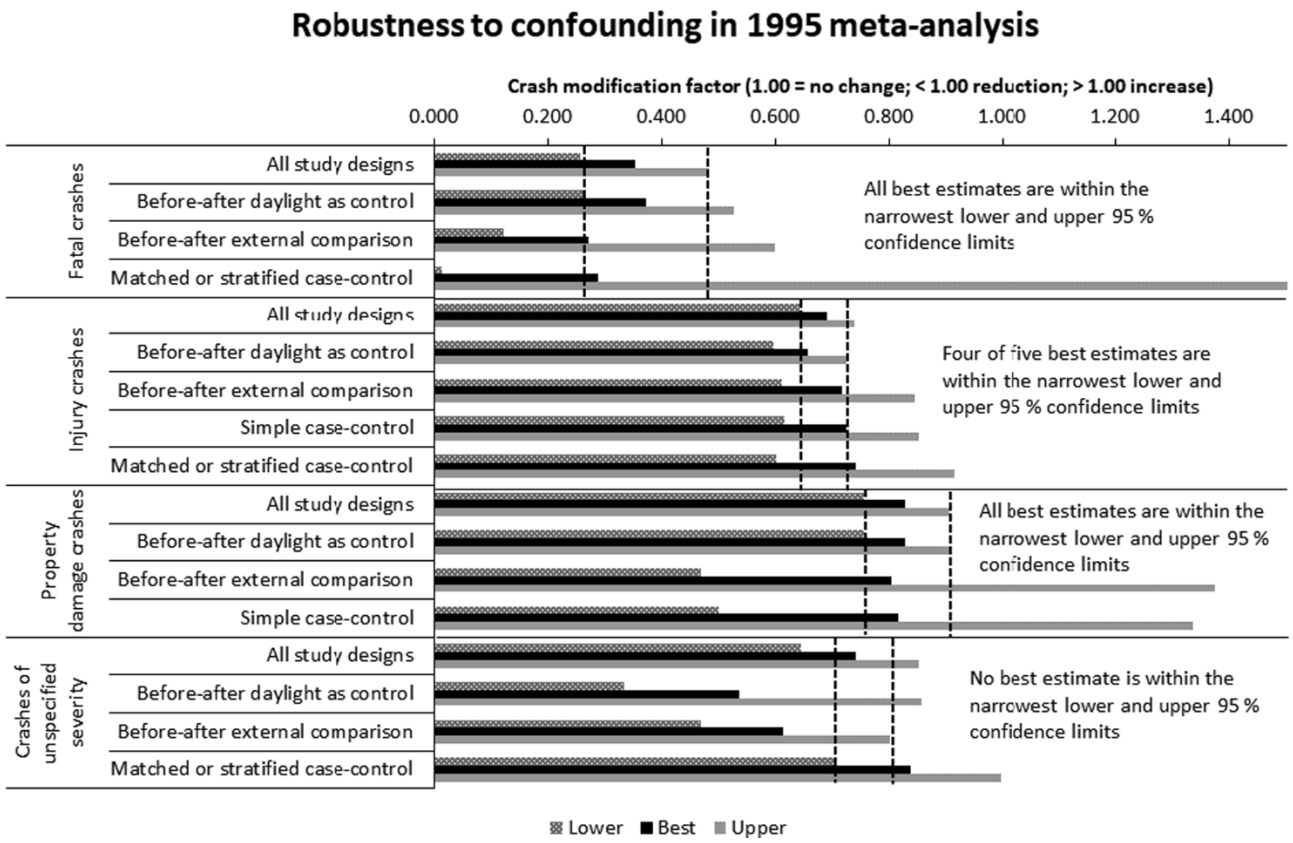

Fig. 1. Robustness to confounding in 1995 meta-analysis.

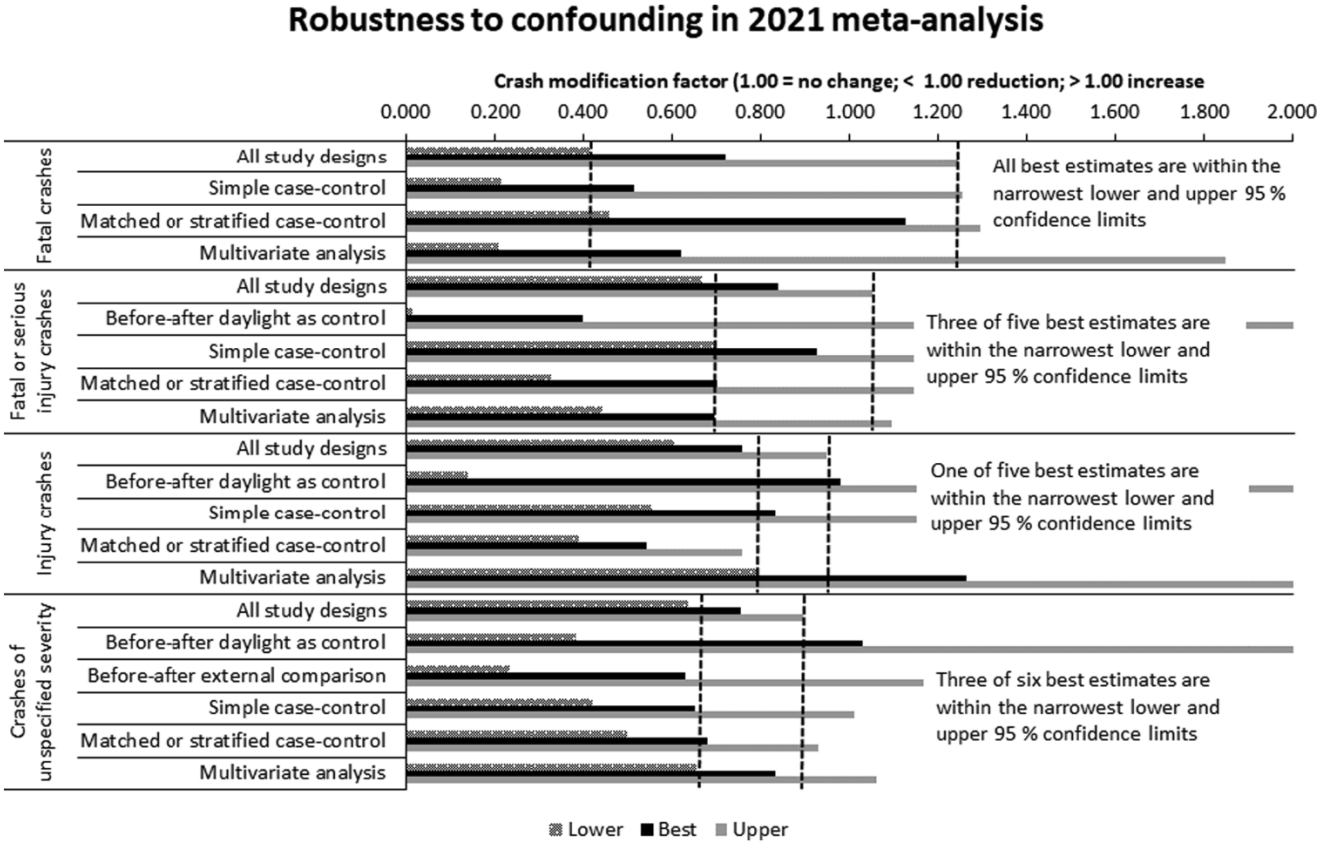

Fig. 2. Robustness to confounding in 2021 meta-analysis.

3. Theories of different ranges

Road safety evaluation research is generally regarded as a having a poorly developed theoretical foundation. There have been attempts at formulating highly general theories about the effects of road safety measures, like the theory of risk homeostasis (Wilde 1982). However, theories at a high level of generality usually cannot guide the interpretation of specific road safety evaluation studies. For that purpose, one needs either what sociologist Robert Merton (1949) referred to as "theories of the middle range", or perhaps even theories of a small range.

There are few examples of theories of the middle range that may guide the interpretation of road safety evaluation studies. Perhaps the frameworks proposed by Evans (1985) and Elvik (2004) may serve as examples. These frameworks model the effects of road safety measures in terms of an "engineering effect" and a "behavioural adaptation effect". Evans tried to quantify behavioural adaptation in terms of a human feedback parameter. These models are not developed to the level of full theories (i.e. specific testable hypotheses about the effects of road safety measures) but remain conceptual schemes. They do not predict, at least not in quantitative terms, the effects of specific road safety measures.

Given the fact that the effects of many road safety measures vary, what is needed are theories, perhaps specifically adapted to each road safety measure, about the sources of variation in their effects. Such theories can be termed low-level theories (Noyes et al. 2016), because they refer only to a single road safety measure. The objective of low-level theories of the effects of road safety measures is to identify sources of systematic variation in effects and propose hypotheses about the direction, and in some cases also the size, of these variations. If the hypotheses about variation in effects are supported, this shows that the results of evaluation studies are likely to at least mostly reflect the effects of the road safety measure, rather than confounding factors evaluation studies did not control for. In other words, support for a low-level theory lends theoretical validity to knowledge and indicates that low internal validity may not be an issue, or at least not a sufficient reason for rejecting the results of evaluation studies.

4. The case of road lighting

4.1. Hypotheses about variation in the effects of road lighting

A low-level theory about variation in the effects of road lighting is proposed in Table 2. A set of generative assumptions are made. From these, hypotheses are derived. Based on the hypotheses, the expected pattern of findings in evaluation studies is specified.

The generative assumptions are intended as statements of well-established truths. In this paper, no attempt has been made to support the generative assumptions by citing original studies; a fairly comprehensive and basic treatment of the scientific foundations of road lighting is given by Ketvirtis (1977). The hypotheses state the expected variation in the effects of road lighting, e.g. with respect to ambient light level (twilight versus darkness), crash severity, groups of road users or types of traffic environment. Finally, the results that support the hypotheses are specified.

A critic might argue that it is always possible to formulate such hypotheses after looking at the results of evaluation studies and framing the hypotheses so that they are all supported. This will give an impression of high theoretical validity, when in fact it is nothing more than a post-hoc rationalisation of study findings, which can be twisted whenever new findings suggest new interpretations. There may often be an element of truth in this criticism. However, hypotheses formulated so as to conform to the pattern of results at a given time, can always be rejected, or reformulated by subsequent studies. Future tests of hypotheses are always possible. The hypotheses proposed here about the effects of road lighting follow directly from the generative assumptions and are not based on the findings of evaluation studies. Ten hypotheses

Table 2
Hypotheses about variation in the effects of road lighting.

Generative assumptions	Hypotheses derived from assumptions	Predicted observations if hypotheses are true
A1: Increasing the distance at which objects or road users can be detected increases the possibility of avoiding crashes	H1 (from A1 and A2): Road lighting reduces the expected number of crashes in hours of darkness	A reduction of the number of crashes in darkness will be found when road lighting is installed; the number of crashes in daylight will not be reduced
A2: Road lighting increases detection distances (in metres) in darkness		
A3: The increase in detection distances associated with road lighting is largest at the lowest levels of atmospheric light	H2 (from A2 and A3): Road lighting will have a larger effect in darkness than in twilight	The percentage reduction in crashes associated with road lighting will be larger in darkness than in twilight
A4: Forests or mountains can create locations that are darker than normal		
A5: There will be systematic variation in the share of crashes occurring in darkness	H3 (from A4 and A5): Road lighting will have a larger effect when the share of crashes in darkness is high than when it is low	The percentage reduction in crashes in darkness will be greater the higher the share of crashes in darkness is
A6: An increased detection distance allows road users more time to brake or manoeuvre; this reduces kinetic energy	H4 (from A6): Road lighting reduces the severity of crashes in darkness	There will be a larger percentage reduction in severe crashes than in less severe crashes
A7: The increase of detection distance produced by road lighting is largest for road users that are small and/or do not use lights or reflective devices	H5 (from A4): Road lighting will have a larger effect for pedestrians and cyclists than for motor vehicles	There will be a larger percentage reduction in crashes involving pedestrians or cyclists than in crashes involving motor vehicles only
A8: The increase in detection distance produced by road lighting is larger the more intense lighting is	H6 (from A5): Intense road lighting will have a larger effect on crashes than dim road lighting	There will be larger percentage reduction in crashes for high-intensity road lighting than for low-intensity road lighting
A9: Dimming road lighting reduces its contribution to increasing detection distance	H7 (from A9): Reducing the intensity of road lighting from its previous level is associated with an increase in crashes in darkness	There will be an increase in crashes in darkness when road lighting is switched off fully or partly or its intensity reduced
A10: Variation in the intensity of lighting may create dark spots or sections in which objects are more difficult to detect	H8 (from A10): Road lighting of high uniformity will have a larger effect than road lighting of low uniformity	There will be larger percentage reduction in crashes the higher the uniformity of lighting is
A11: Traffic environments vary with respect to the presence of road users who are difficult to detect in darkness	H9 (from A11): Road lighting will have a smaller effect on motorways (where there are no pedestrians or cyclists) than in other traffic environments	The percentage reduction of crashes in darkness associated with road lighting will be smaller on motorways than in other traffic environments
A12: The presence of other sources of artificial lighting reduces the increase in risk during darkness	H10 (from A12): Road lighting will have a smaller effect on crashes in built-up areas than in non-built-up areas	The percentage reduction in crashes in darkness associated with road lighting will be smaller in built-up areas than in other traffic environments

are proposed in Table 2. The next section explains how the hypotheses were tested.

4.2. Testing the hypotheses

The hypotheses have been tested mainly by relying on two meta-analyses of studies evaluating the effects on crashes of road lighting. These meta-analyses were reported twenty-six years apart. The first, by Elvik (1995), included 37 studies. These studies are listed in part B of the references. The second, by Høye (2021), included 35 studies. These studies are listed in part C of the references.

The two meta-analyses are based on completely non-overlapping primary studies. They therefore provide two independent tests of the hypotheses, based on different studies. Both meta-analyses were based on systematic searches for relevant studies. This, of course, does not guarantee that every study is found, and publication bias (Rothstein et al. 2005) remains a concern in every meta-analysis. The first analysis (Elvik 1995) assessed the potential presence of publication bias by visual inspection of funnel plots. None of them indicated publication bias. Subsequently, better methods for assessing the potential presence of publication bias have been developed, in particular the trim-and-fill method (Duval and Tweedie 2000A, 2000B; Duval 2005). This method is now routinely applied in all meta-analyses the authors perform. In the most recent meta-analysis (Høye 2021), there was no evidence of publication bias.

For testing hypothesis 6, about intense road lighting having larger effect than less intense road lighting, studies that have modelled the relationship between lighting level and road safety were reviewed and summarised. These studies are listed in part D of the references. Two of these studies (Yang et al. 2019) were used to test hypothesis 8, about the effects of uniformity in lighting level.

Twelve functional relationships between lighting intensity and the share of crashes occurring in darkness were extracted from the studies listed in part D of the references. These relationships are shown in Fig. 3. Each curve was fitted by least-squares regression and is the best fitting curve when the following functional forms were compared: linear, logarithmic, quadratic, exponential, power.

Goodness of fit was assessed in terms of R-squared. There was considerable variation between the curves with respect to how well they fitted the data points in the studies. The best fitting curve had an R-squared value of 0.997. The poorest fitting curve had an R-squared value of 0.106.

It is seen that the curves diverge a lot but can be divided into two clusters with respect to the share of crashes in darkness. The upper cluster consists of curves that show how increasing the intensity of road lighting influences crashes when the initial share of crashes in darkness is about 0.50 (50 %). The studies included in this cluster had initial shares of crashes in darkness of 0.436; 0.449; 0.440; 0.549 and 0.466. The weighted mean share was 0.474 with a standard error of 0.042. The lower cluster consists of curves that show how increasing the intensity of lighting influences crashes when the initial share of crashes in darkness is about 0.25 (25 %). The studies in this cluster had initial shares of crashes in darkness of 0.292; 0.336; 0.223; 0.227; 0.268; 0.204 and 0.254. The weighted mean was 0.266 with a standard error of 0.219. The curves in each cluster were combined by assigning a weight to each curve which was inversely proportional to its residual variance:

$$\text{Weight assigned to each curve} = \frac{1}{1 - R^2}$$

This ensured that curves fitting the data closely had a larger weight than those fitting the data poorly. The resulting summary curves are shown in Fig. 4.

A problem when trying to compare and combine the results of studies of the intensity of lighting, is that this has not always been measured the same way. The intensity of lighting can be measured either by the amount of light emitted by the source of lighting (illuminance), by the amount of light reflected from a surface which is illuminated (luminance), or by small target visibility (Fotios and Gibbons 2018). Unfortunately, there is no simple way of converting, for example, illuminance to luminance. It depends, for example, on whether the road surface is dark or consists of lighter material. It depends on whether the road surface is dry or wet. It depends on the height of the source of lighting above the surface that reflects the light hitting it. Therefore, the lowest level of lighting in each study was given the value of 1 and higher

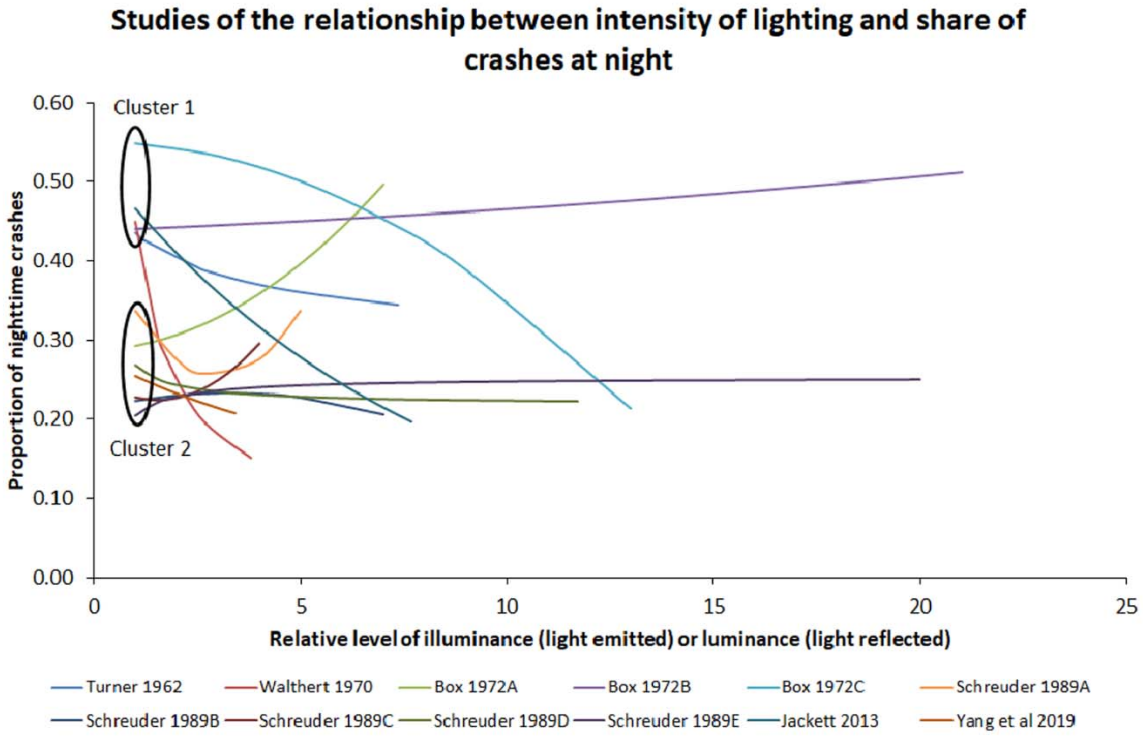

Fig. 3. Studies of the relationship between intensity of lighting and share of crashes at night.

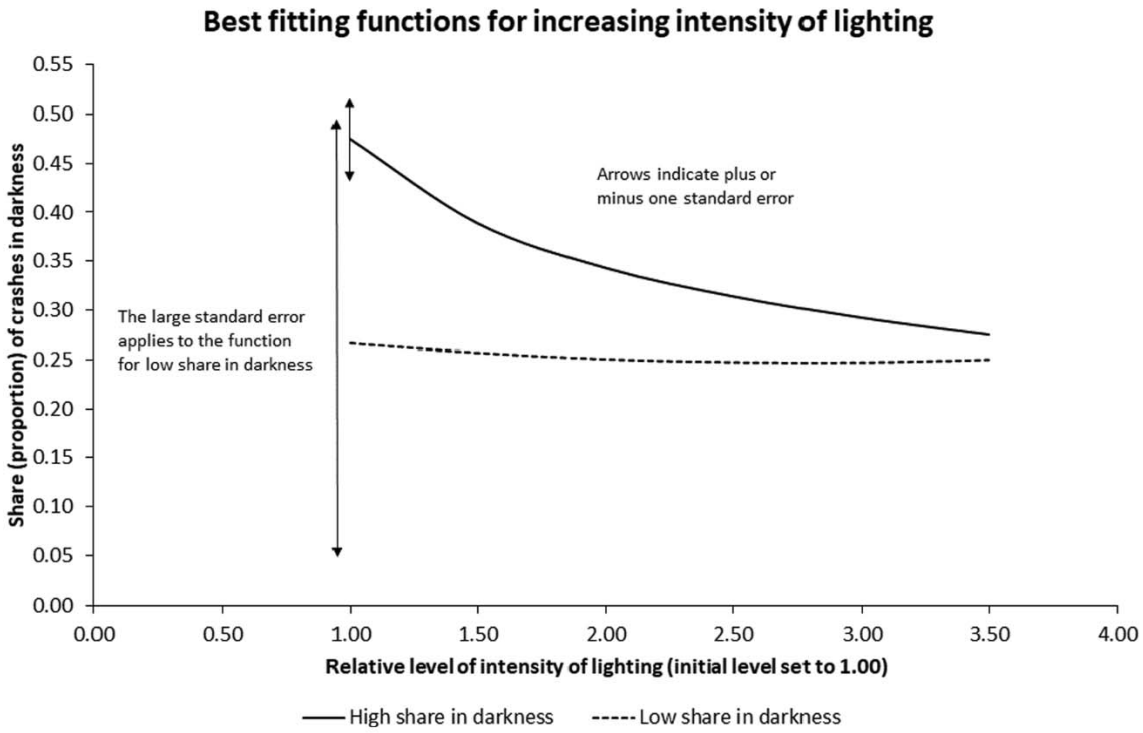

Fig. 4. Best fitting functions for increased intensity of lighting.

intensities given values of 2, 3, etc. The summary curves could only be developed up to a lighting intensity of 3.5 times the initial level. These intensity values are a mixture of illuminance (light emitted) and luminance (light reflected). It is seen that when the initial share of crashes in darkness was high, increasing the intensity of road lighting was associated with a reduction of the share of crashes occurring in darkness. When the initial share of crashes in darkness was low, increasing the intensity of lighting was not associated with a reduction of the share of crashes occurring in darkness.

Table 3 summarises the results of the tests of the hypotheses. Four hypotheses (1, 4, 5, 8) were supported by the findings of more than one study. One hypothesis (2) was supported by the findings of a single study. Two hypotheses (6, 7) were partly supported, i.e. some but not all findings supported them. One hypothesis (3) was partly supported by comparing the findings of before-and-after studies to those of cross-sectional studies. Finally, two hypotheses (9, 10) did not get clear support.

The lack of clear support for hypotheses 9 and 10 is perhaps not surprising, as the generative assumptions are somewhat ambiguous and perhaps incomplete for these hypotheses. Hypotheses 9 was justified by referring to the assumption that road users who are particularly difficult to detect in the dark, i.e. pedestrians and cyclists, are not found on motorways. The absence of this risk factor was taken to imply a smaller effect on motorways than on other roads. Against this, one might argue that speeds are higher on motorways, and that this entails a smaller safety margin than lower speeds.

As far as the urban traffic environment is concerned, the assumption was made that there are more sources of artificial lighting in urban areas than in rural areas (more buildings that are lit up in various ways), so that the added light provided by road lighting makes less difference. Against this, one could argue that there are more pedestrians and cyclists in urban areas than in rural areas. If this assumption is added, the implications of the generative assumptions become less clear.

5. Discussion

Most road safety evaluation studies are observational and subject to many threats to validity. When interpreting the results of such studies, one should always consider all aspects of validity. The validity framework of Shadish, Cook and Campbell (2002) is useful in helping researchers think of all things that can go wrong in a study and make it impossible to interpret the study substantively, i.e. as showing real effects.

It is reasonable to regard most studies evaluating the effects on crashes of road lighting as having low statistical conclusion validity and low internal validity. Replication can be viewed as a cure for low statistical conclusion validity. Although each study may be based on a small convenience sample, if there are many studies, made in different contexts, and their findings are consistent, we may be entitled to conclude that the findings do not merely reflect local circumstances, but are valid across these local circumstances. In short: although each local context is unique, successful replication shows that this does not matter and refutes the argument that all knowledge is local.

Low internal validity may not be a major problem if the results of evaluation studies are robust with respect to confounding. To be robust with respect to confounding means that poor control of confounding factors, which characterises many evaluations of road lighting, does not influence results very much. Robustness to confounding was assessed by comparing summary estimates of effect based on study designs with different degrees of control for confounding. It was found that the results of studies using designs that differ with respect to their control for confounding factors were close, suggesting that they are robust with respect to confounding.

The main argument made in this paper is that if the results of all studies that have evaluated the effects of road lighting make sense from a theoretical point of view, then one is more justified in believing the results of these studies than in rejecting them. To help in such an assessment, a low-level theory about variation in the effects of road lighting was proposed. The term low-level refers to the fact that the hypotheses proposed apply to road lighting only and not to other road safety measures. Ten hypotheses were proposed. Most of them were supported. Only two hypotheses were not supported by the findings of evaluation studies. On the whole, this suggests that evaluation studies have been able to measure the effects of road lighting and do not merely

Table 3
Results of tests of hypotheses about variation in the effects of road lighting.

Hypothesis	Study	Conditions	Percentage change in number of crashes (95 % confidence interval)		Comments
H1: Reduction of crashes in darkness	Elvik, 1995 (top)Høye 2021 (bottom)	All	–23 (−25; −20) −21 (−29; −13)		Hypothesis is supported
H2: Larger reduction in darkness than in twilight	Wanvik 2009B	Darkness Twilight	−46 (−50; −42) −31 (−36; −26)		Hypothesis is supported, but only this study was found
H3: Larger reduction when high share of crashes in darkness	Elvik 1995	Share in darkness High (>50 %)Medium (33–50 %) Low (<33 %)	Before-after −35 (−41; −28) −21 (−25: −17) –22 (−26; −17)	Cross-sectional –32 (−39; −24) −15 (−21; −7) −21 (−27; −14)	Hypothesis is partly supported; pattern in before-after studies could be the result of regression-to-the-mean; this is not the case in cross-sectional studies
H4: Larger reduction in serious crashes than in less serious crashes	Elvik 1995 (left)	Fatal crashes	−65 (−75; −52)	−49 (−63; −30)	Hypothesis is supported
	Høye 2021 (right)	Injury crashes	−29 (–32; −26)		
		Property-damage only	−17 (−21; −13)	−21 (−40; +2) −10 (−35; +24)	
H5: Larger reduction for pedestrians and cyclists than for motor vehicles only	Elvik 1995 (left)	Pedestrians	−52 (−58; −45)	−45 (−62; −19)	Hypothesis is supported
	Høye 2021 (right)	Cyclists			
		Motor vehicles only	−17 (−21; −13)	−60 (−65; −54) −8 (−19; +4)	
H6: Larger reduction for high-intensity than low-intensity lighting	See studies listed in Fig. 3	Increasing intensity up to 3.5 times baseline	−42 (−34; −50) if high share in darkness		Hypothesis is supported if the share of crashes in darkness is high; otherwise not
			−6 (−49; +37) if low share in darkness		
H7: Increase in crashes when lighting is switched off/ reduced	Elvik and Vaa 2004 (top)	Most commonly half the lamps are switched off	+17 (+9; +25) injury crashes		Hypothesis is supported, except for one result referring to injury crashes (based on Monsere and Fischer 2008)
	Høye 2021 (bottom)		+27 (+9; +50) property damage crashes +9 (+0; +18) severity not specified −11 (–22; +2) injury crashes		
H8: Larger reduction with high uniformity of lighting levels	Jackett, Frith 2013 (top)	Going from minimum to maximum uniformity	−8 (uncertainty not stated)		Hypothesis is supported
	Yang et al. 2019 (bottom)		−7 (−11; −3)		
9: Smaller reduction on motorways than elsewhere	Elvik 1995 (left) Høye 2021 (right)	Motorways All other roads	–23 (−25; −20) –23 (−25; −20)	−14 (−40; +23) –22 (−30; −13)	There is only a weak tendency in the direction predicted by the hypothesis
H10: Smaller reduction in urban areas than elsewhere	Elvik 1995 (left) Høye 2021 (right)	Urban areas All other areas	–22 (−25; −19) –23 (−25; −21)	−18 (−36; +5) –22 (−31; −12)	There is only a weak tendency in the direction predicted by the hypothesis

reflect confounding factors that were not controlled for.

6. Conclusions

The main conclusions of the study presented in this paper can be summarised as follows:

1. The interpretation of road safety evaluation studies can be guided by developing low-level theory identifying sources of variation in the findings of evaluation studies.
2. If a systematic pattern in study findings is found, supporting the theory, it is more likely that studies mainly show the effects of the road safety measure rather than mainly the effects of confounding factors not controlled for by evaluation studies.
3. A case of using low-level theory was developed for road lighting. Ten hypotheses were proposed and most of them were supported.
4. Taken as a whole, the research literature shows effects it is more reasonable to attribute to road lighting than to confounding factors.

CRediT authorship contribution statement

Rune Elvik: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Conceptualization. **Alena Katharina Høye:** Formal analysis, Writing – review & editing, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The study reported in this paper benefitted from the funding given to the Handbook of Road Safety Measures during a long period by the Norwegian Ministry of Transport and the Norwegian Public Roads Administration, grant number 1175.

References

Duval, S., 2005. The trim and fill method. In: Rothstein, H., Sutton, A.J., Borenstein, M. (Eds.), Publication bias in meta-analysis: Prevention, assessment and adjustments, 127-144. Chichester, John Wiley and Sons. Part A. General References.
Duval, S., Tweedie, R., 2000A. Trim and fill: a simple funnel plot based method of testing and adjusting for publication bias in meta-analysis. J. Am. Statist. Assoc. 95, 89-98.
Duval, S., Tweedie, R., 2000B. A non-parametric trim and fill method of assessing publication bias in meta-analysis. Biometrics 56, 455–463.
Elvik, R., 1995. Meta-analysis of evaluations of public lighting as accident countermeasure. Transport. Res. Rec. 1485, 112–123.
Elvik, R., 1997. Evaluation of road accident blackspot treatment: a case of the Iron Law of evaluation studies? Acc. Anal. Prevent. 29, 191–199.

Elvik, R., 2004. To what extent can theory account for the findings of road safety evaluation studies? Acc. Anal. Prevent. 36, 841–849.
Elvik, R., 2011. Assessing causality in multivariate accident models. Accid. Anal. Prevent. 43, 253–264.
Elvik, R., 2012. The range of replications technique for assessing the external validity of road safety evaluation studies. Acc. Anal. Prevent. 45, 272–280.
Elvik, R., 2021. Why are there so few experimental road safety evaluation studies: could their findings explain it? Acc. Anal. Prevent. 163, 106467.
Elvik, R., Vaa, T., 2004. The Handbook of Road Safety Measures. Elsevier, Oxford.
Evans, L., 1985. Human behavior feedback and traffic safety. Human Factors 27, 555–576.
Hauer, E., Bonneson, J.A., Council, F., Srinivasan, R., Zegeer, C., 2012. Crash modification factors. Foundational issues. Transport. Res. Rec. 2279, 67–74.
Høye, A.K., 2021. Revisjon av Trafikksikkerhetshåndboken. 1.18 Vegbelysning. Arbeidsdokument 8.2.2021. Oslo, Transportøkonomisk institutt.
Ketvirtis, A., 1977. Road Illumination and Traffic Safety. Prepared for Road and Motor Vehicle Traffic Safety Branch, Transport Canada. Ottawa, Transport Canada.
Merton, R.K., 1949. On sociological theories of the middle range. In: In Merton, R.K. (Ed.), Social Theory and Social Structure. The Free Press, New York, pp. 39–53.
Monsere, C.M., Fischer, E.L., 2008. Safety effects of reducing freeway illumination for energy conservation. Acc. Anal. Prevent. 40, 1773–1780.
Noyes, J., Hendry, M., Booth, A., Chandler, J., Lewin, S., Glenton, C., Garside, R., 2016. Current use was established and Cochrane guidance on selection of social theories for systematic reviews of complex interventions was developed. J. Clin. Epidemiol. 75, 78–92.
Fotios, S., Gibbons, R., 2018. Road lighting research for drivers and pedestrians: the basis of luminance and illuminance recommendations. Light. Res. Technol. 50, 154–186.
Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin, Boston.
Wilde, G.J.S., 1982. The theory of risk homeostasis: implications for safety and health. Risk Analysis, 2, 209-225. Part B. Studies included in 1995-meta-analysis.
Wanvik, P.O., 2009B. Effects of road lighting: An analysis based on Dutch accident statistics 1987–2006. Acc. Anal. Prevent. 41, 123-128. Part D. Studies of the relationship between lighting level and road safety.
Yang, R., Wang, Z., Lin, P.-S., Li, X., Chen, Y., Hsu, P.P., Henry, A., 2019. Safety effects of street lighting on roadway segments: development of a crash modification function. Traffic Inj. Prevent. 20, 296–302.