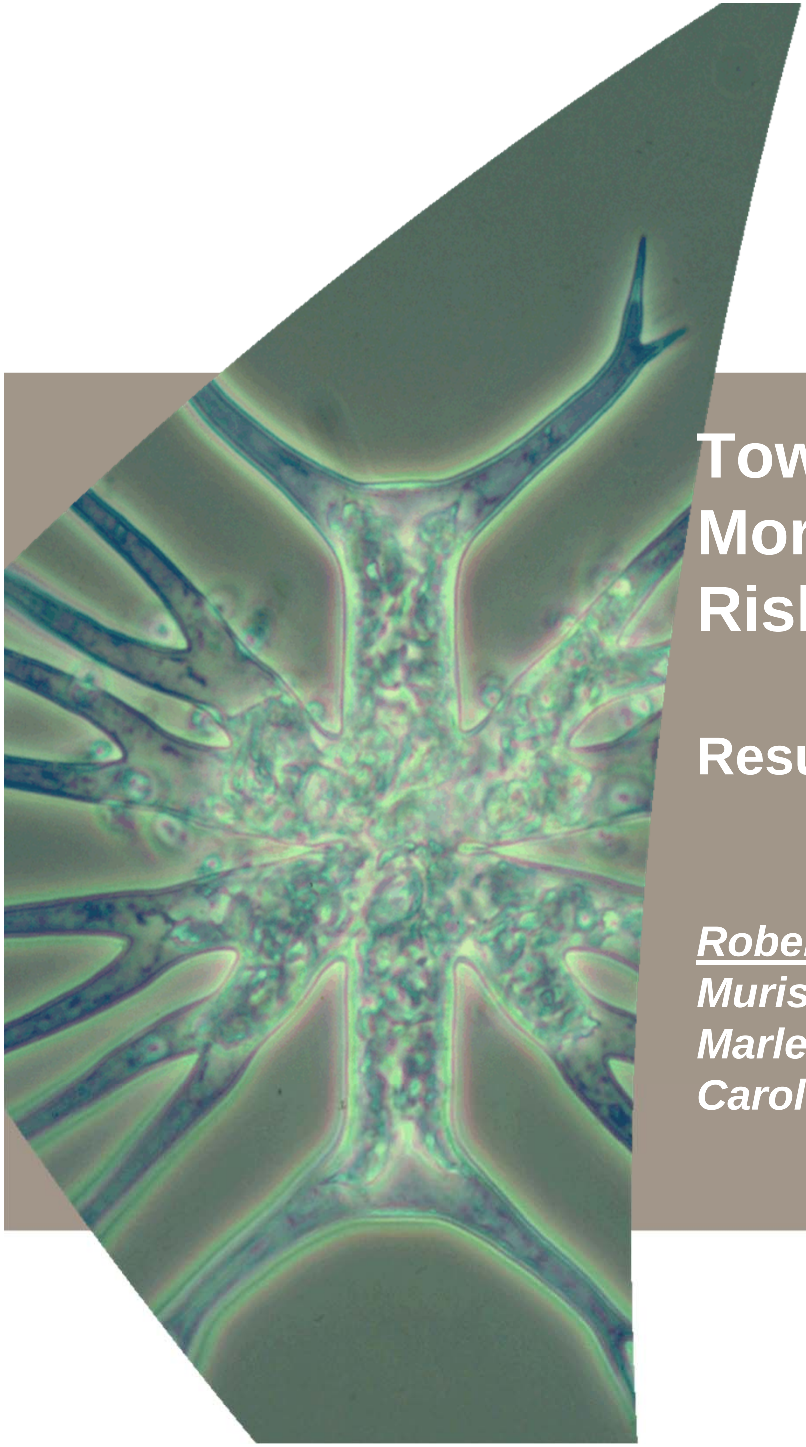


oekotoxzentrum
centre ecotox

Schweizerisches Zentrum für angewandte Ökotoxikologie
Centre Suisse d'écotoxigologie appliquée
Eawag-EPFL

Towards Klimisch 2.0? – More Transparency and Quality in Risk Assessment

Results of the Klimisch ringtest phase I + II

Robert Kase *(Robert.Kase@oekotoxzentrum.ch)*
Muris Korkaric (Muris.Korkaric@eawag.ch)
Marlene Agerstrand (Marlene.Agerstrand@itm.su.se)
Caroline Moermond (Caroline.Moermond@rivm.nl)

Multilateral group 17th May, including points of discussion from SETAC PAG and ERAAG 2013

Structure

- Background and aims
- Participation
- Results of phase I+II
  - Relevance assessment comparison
  - Reliability assessment comparison
  - Consistency evaluation
- Assessment of the two systems
- Preliminary conclusions
  - Regarding: transparency , strictness, consistency, quality gain and applicability
- Discussion & Outlook

Background

- The Klimisch code evaluation system is commonly used for study assessments in different regulations, e.g. REACh 2006, TGD for EQS 2011, EM(E)A 2006, 91414/EEC and 1107/2009/EC. However, recent studies indicated the need for an updated evaluation system. **Therefore, we performed a ringtest to compare the current Klimisch evaluation scheme with an updated checklist version**.

- **In total, 8 aquatic ecotox studies were assessed** including **different taxonomic groups,** tested **substance classes** (hormonally active substances, industrial chemicals, biocides and pesticides and pharmaceuticals) and **quality levels**
  - **Phase I:** evaluation with the current **Klimisch** evaluation (end of 2012)
  - **Phase II:** evaluation with **checklist system** (spring 2013).

- Now we are able to compare the **functionality** of the **current Klimisch evaluation system (phase I)** **in comparison** to an updated **checklist evaluation approach (phase II)** on a statistical basis.

Aims and motivation

Our overall aims are to:

- **identify the current weaknesses** in environmental hazard assessments
- to improve the **safety and transparency** of the current assessment system
- to **give improved guidance** to which information is necessary **to allow a regulatory use of scientific studies**
- to help **avoiding endless discussions** about the reliability, relevance and plausibility of single studies.

Many cooperation partners participated in this activity, indicating an overarching interest in this issue

Participation in phase I + II

Around 80 international participants from 35 groups/organizations:

- **regulation:**
  CAN, DK, GER, FR, NL, SWE, UK, USA, EU (ECHA) and working groups

- **industry and stakeholder organizations**:
  Bayer (GER), BASF (GER), Givaudin International SA (CH), Golder Associates Inc.(USA), ECETOC (EU), Harlan (CH), Monsanto Europe (B), Pfizer (USA)

- **science, advisory and assessment institutions**:
  Astrazeneca (UK), CEFAS (UK), CEHTRA (FR), CERI (J), Deltares (NL), DHI (Singapoore), ECT (GER), Eurofins AG (CH), GAB Consult (GER), ITEM (GER), RIVM (also regulatory institution), RWS (NL), SETAC Pharmaceutical Advisory Group, SETAC Global Ecological Risk Assessment Group, Swiss Centre for Applied Ecotoxicology (CH), TSGE (UK), wca (UK)

Thank you for your participation, your contributions, and your invested work.

Experience level of participants in study assessment

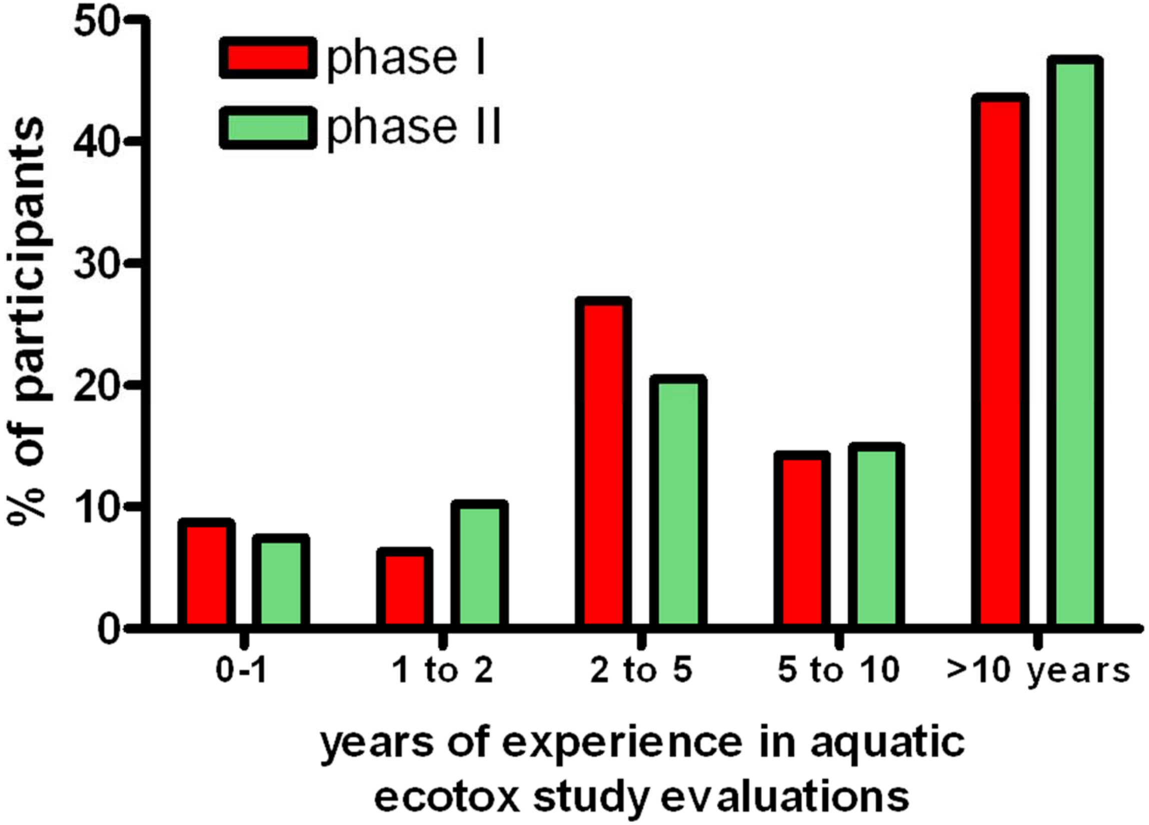


Fig. 1: Experience level of participants

- Most participants had more than 10 years experience in study evaluations
- In both phases a relatively experienced group of assessors was contributing

	phase I	phase II
mean experience	> 6.5 years	> 6.7 years
participation level	78%	69%
# questionnaires	126	107

Relevance evaluation in phase I

STUDY	A	B	C	D	E	F	G	H	MEAN I	MEAN II
n	13	12	19	16	11	13	20	19	15.37	13.37
MEAN	1.54	1.83	1.79	1.88	1.36	1.62	1.90	1.89	**1.73**	**1.64**

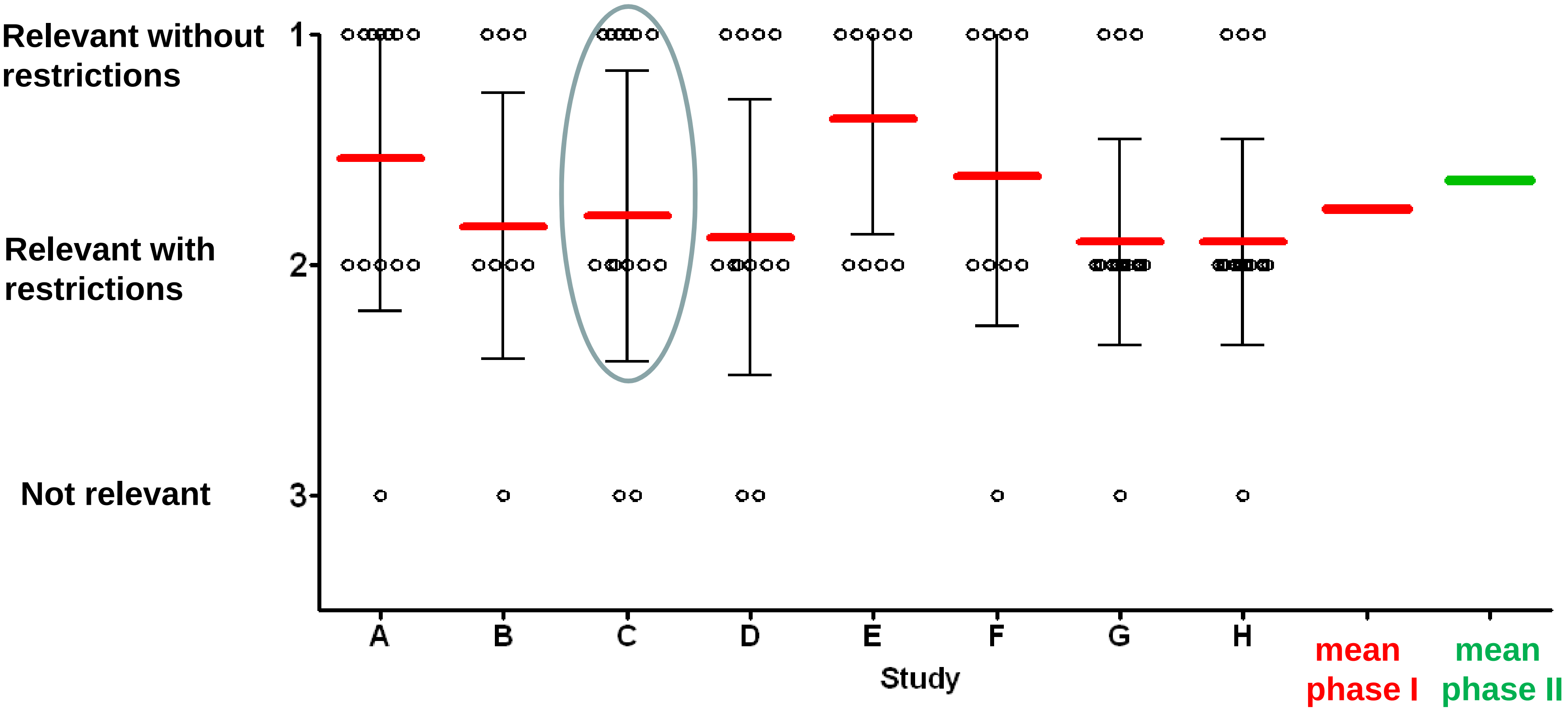
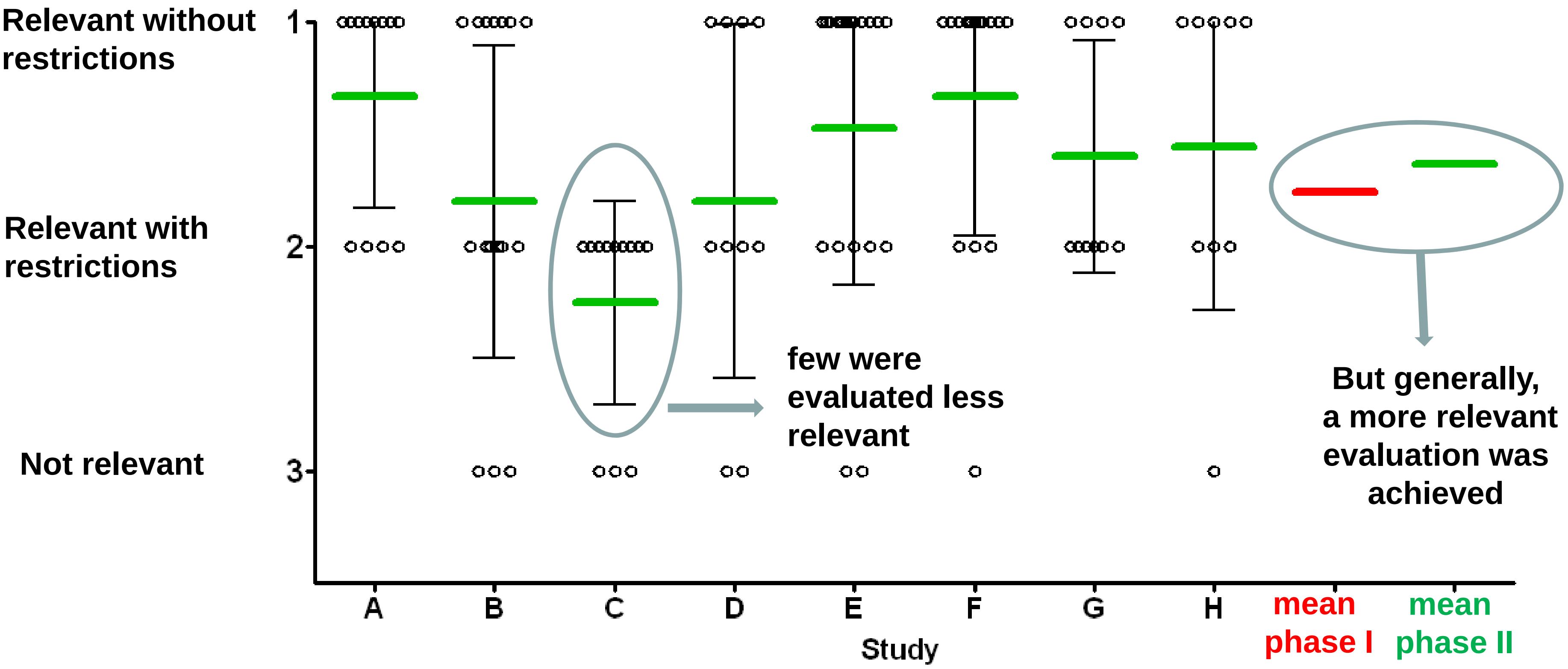


Fig. 2: Relevance assessment of 8 studies in phase I

Relevance evaluation in phase II

STUDY	A	B	C	D	E	F	G	H	MEAN I	MEAN II
n	12	20	12	10	19	15	10	9	15.37	13.37
MEAN	1.33	1.80	2.25	1.80	1.47	1.33	1.6	1.56	**1.73**	**1.64**



few were evaluated less relevant

But generally, a more relevant evaluation was achieved

Fig. 3: Relevance assessment of 8 studies in phase II

Results of study evaluations, regarding relevance

Phase I + phase II

- **Both** relevance evaluations in **phase I + phase II** are **relatively similar**

- **Most** studies were evaluated **between R1 «relevant without restrictions» and R2 «relevant with restrictions»** (**mean I = 1.73 vs mean II = 1.64**)

- **Only few studies were evaluated as not relevant (%R3= 7.3% vs %R3= 11.2%) and 90% of evaluations indicate usable R1 or R2 studies**

- With **both approaches** there is a probability that different assessors would categorize the same study differently (**inconsistency** mainly **between R1 and R2**)

OUTLOOK (to be discussed): For further investigations with the **checklist approach** we could **weight and score the fulfilled relevance criteria** as a **transparent indicator for the relevance** of a study (see discussion point at the end).

Reliability evaluation in phase I

STUDY	A	B	C	D	E	F	G	H	MEAN I	MEAN II
n	14	12	19	16	11	13	20	20	15.625	13.25
MEAN	2.25	2.45	2.78	2.38	1.45	2.31	2.55	2.71	**2.42**	**2.64**

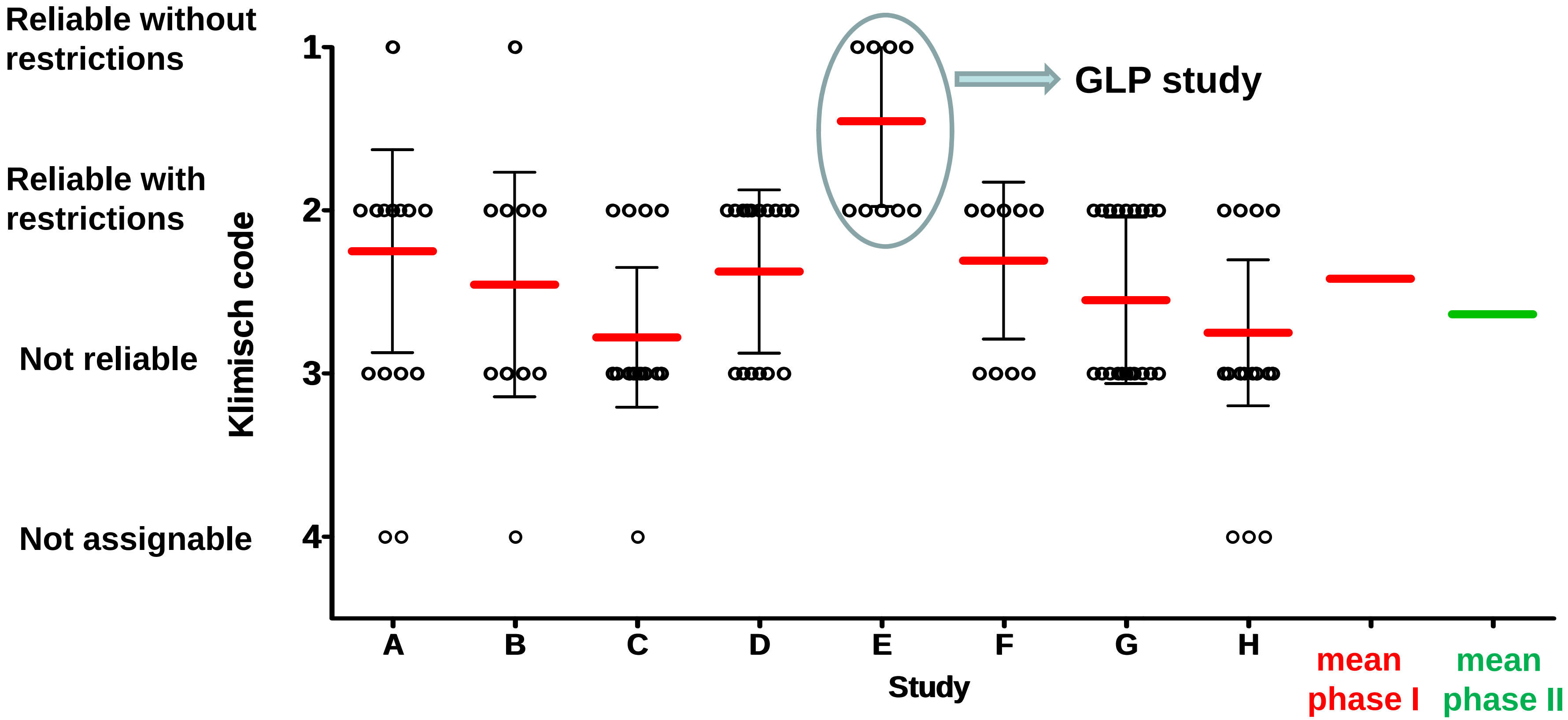
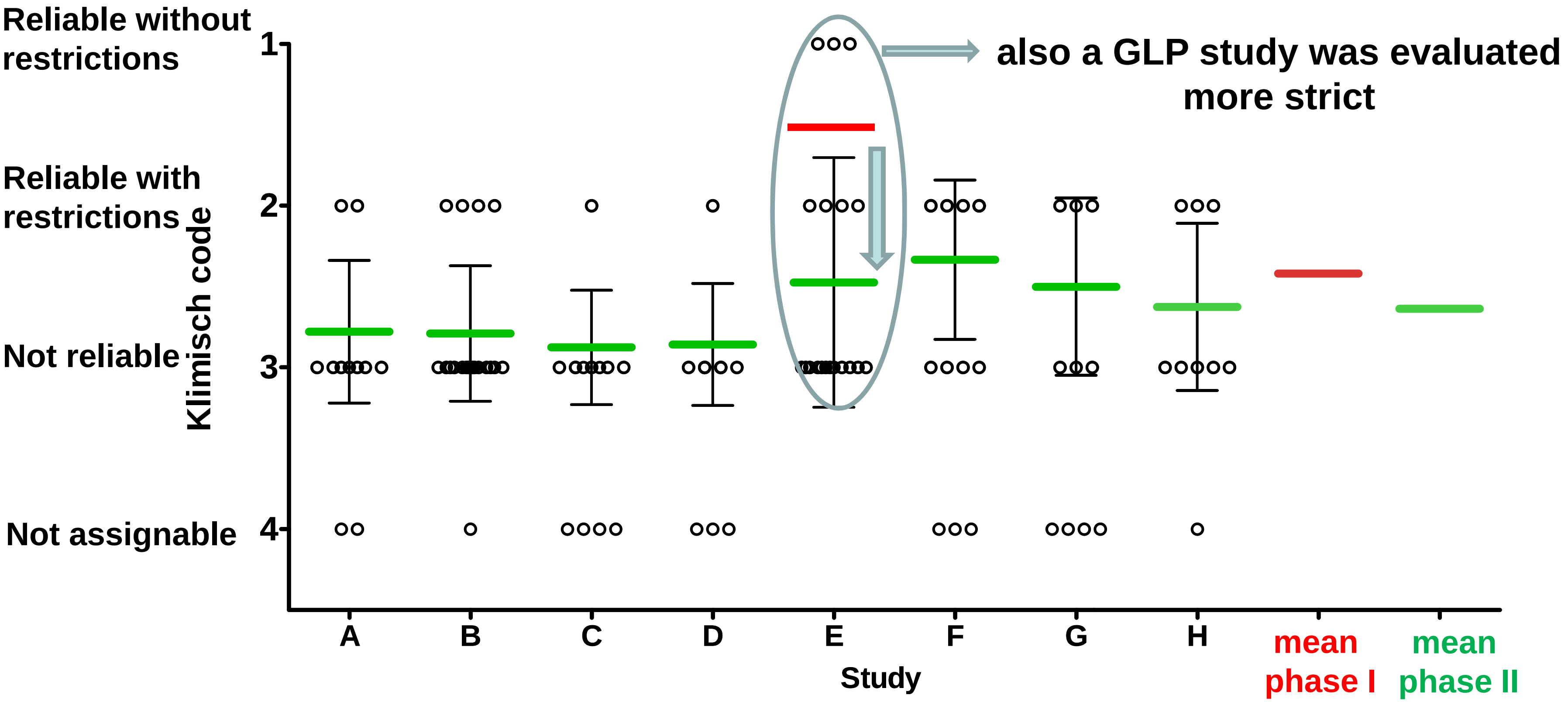


Fig. 4: Phase I reliability evaluation of 8 studies. MEAN and STDEV for Klimisch code 1-3 was calculated, the Klimisch 4 datapoints were shown additionally

Reliability evaluation in phase II

STUDY	A	B	C	D	E	F	G	H	MEAN I	MEAN II
n	11	20	12	10	19	15	10	9	15.625	13.25
MEAN	2.78	2.79	2.88	2.86	2.47	2.33	2.50	2.63	**2.42**	**2.64**



also a GLP study was evaluated more strict

Fig. 5: Phase II reliability evaluation of 8 studies. MEAN and STDEV for Klimisch code 1-3 was calculated, the Klimisch 4 datapoints were shown additionally

Reliability evaluation phase I vs phase II

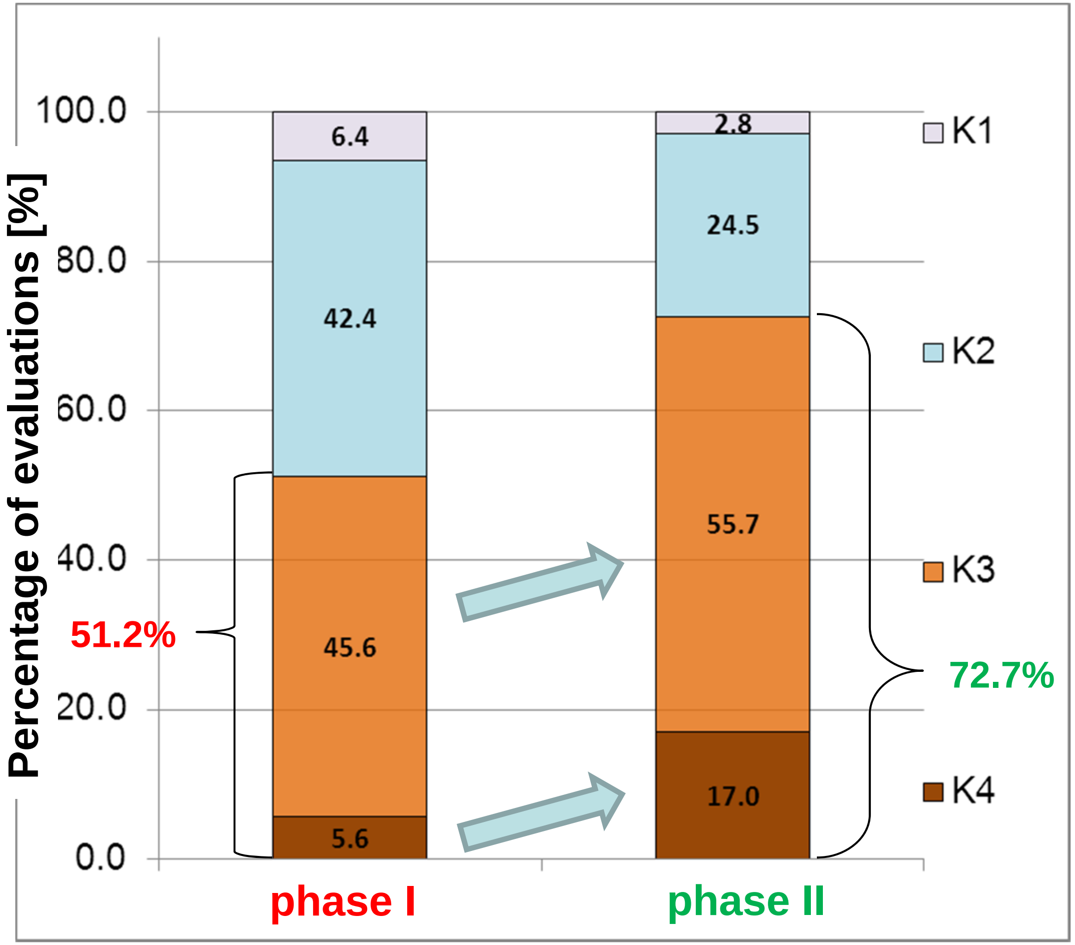


Fig. 6: Klimisch evaluations in phase I+ II

- With the current Klimisch system 51.2% of studies were not directly usable (K3+K4) studies.
- Only 5.6% of studies were evaluated as non assignable (K4)
- With the checklist system 72.7% of studies were not directly usable
- With the checklist system 17% of studies were evaluated as non assignable (K4)
- We intended with the checklist a quality gain, resulting in 21.5% more studies did not fullfill K1 or K2.

The checklist system is around 20% more strict in evaluating the reliability

Results of study evaluations, regarding reliability

Phase I + phase II

- **Both** evaluations in **phase I + phase II are quite different**, the **checklist system is around 20% more strict**

- **Most** of the **study evaluations are between Klimisch 2-3**, but with different quality levels (mean I=2.42 vs mean II=2.64)

- **Many** studies **were evaluated as not usable (%K3+K4= 51.2% vs %K3+K4= 72.7%)**,

⇒ **reliability is an important discriminator**

- **Few studies** were evaluated as **not assignable** in **phase I (only 5.6%)**, and relatively **more in phase II (17%)**

- There is a **relatively high probability** that different assessors would **categorize the same study differently** (**high inconsistency** mainly between reliable with restrictions and not reliable) .

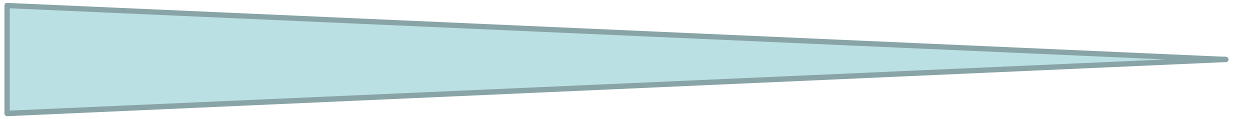
⇒ **This can result in different exclusions of studies and is directly EQS relevant**

OUTLOOK: For further characterization of the checklist approach, we will evaluate the consistency change between phase I+II.

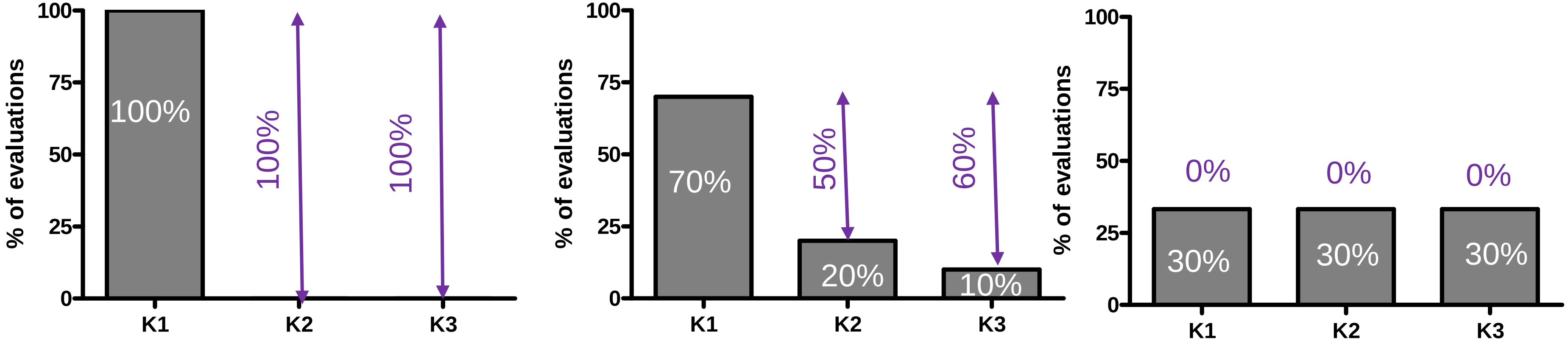
Consistency analysis

Consistency = average distance to strongest class

Maximum consistency



Minimum consistency



Consistency: **100%** **55%** **0%**

Consistency change for reliability evaluation between phase I and II

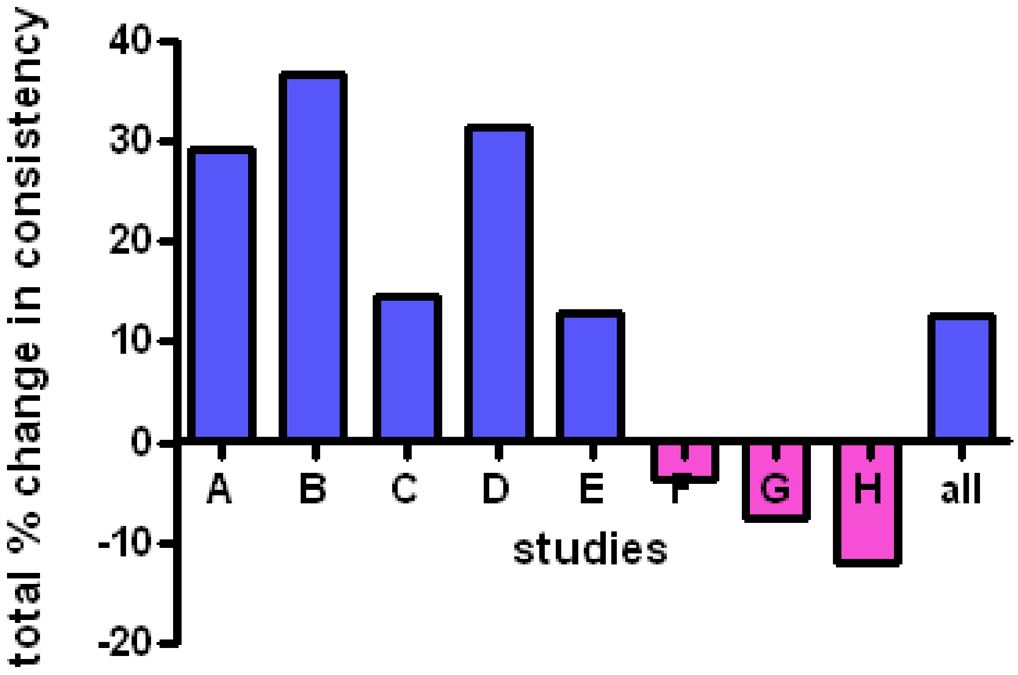


Fig. 7: Consistency change between phase I+ II

- 5 studies were evaluated more consistent in phase II
- 3 studies were evaluated less consistent in phase II
- The consistency gain was stronger in the 5 studies than the loss in the 3 studies

In general a consistency gain was achieved in phase II

Assessment of the two systems during phase I+II

Q1: The evaluation system allows enough **accuracy** for a specific evaluation of **reliability**.

Q2: The evaluation system allows enough **accuracy** for a specific evaluation of **relevance.**

Q3: The evaluation system is **easy** and **applicable** for routine use.

Q4: The use of the evaluation system leads to **consistent results** if the same study is evaluated by different risk assessors.

Q5: The evaluation system depends strongly on **personal expert judgement.**

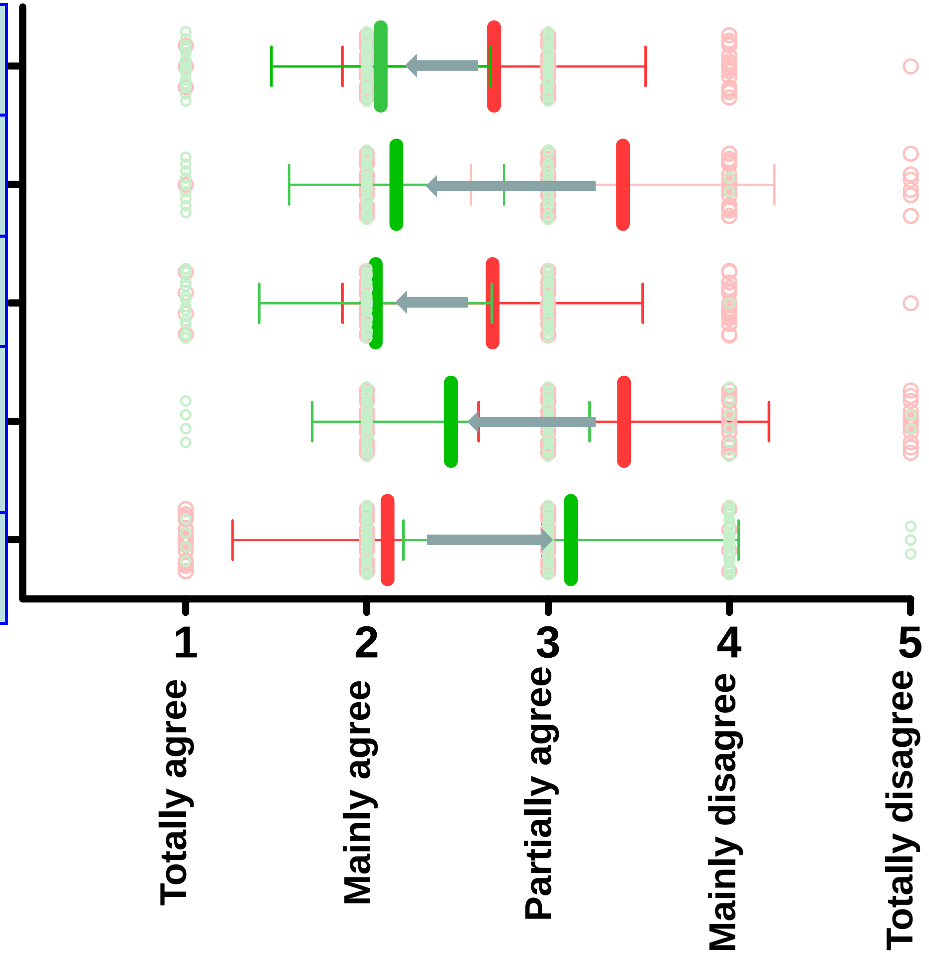


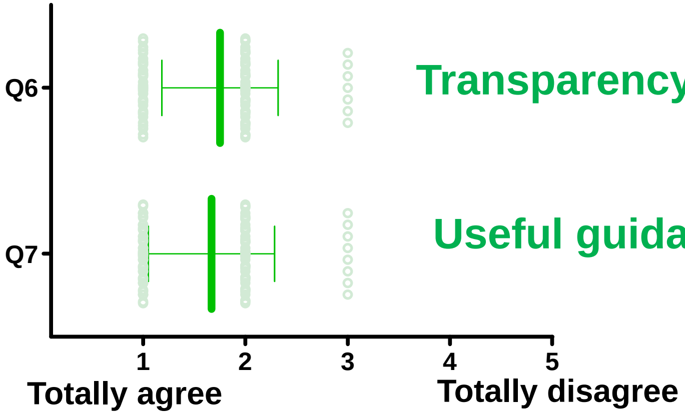
Fig. 8: Assessment of evaluation systems in the questionnaire; red=phase I; green = phase II

Phase I

- **Q1-Q4 were answered not very decided**
- **Q5 The perception of participants is that the current Klimisch evaluation strongly depends on personal expert judgement**

Perception: How was the checklist system evaluated?

Q1: The evaluation system allows enough **accuracy** for a specific evaluation of **reliability**.	Accuracy for reliability evaluation improved
Q2: The evaluation system allows enough **accuracy** for a specific evaluation of **relevance.**	Accuracy for relevance evaluation improved
Q3: The evaluation system is **easy** and **applicable** for routine use.	Applicability improved
Q4: The use of the evaluation system leads to **consistent results** if the same study is evaluated by different risk assessors.	Consistency improved
Q5: The evaluation system depends strongly on **personal expert judgement.**	Personal expert judgement reduced
Q6: The checklist approach **increases the transparency** in comparison to the commonly used Klimisch evaluation.	Transparency increased
Q7:•The **guidance document** to the checklists was **useful** for the study evaluation	Useful guidance



The checklist system was evaluated more positively in all categories

Preliminary conclusions of phase I (Klimisch evaluation system)

- The results of Klimisch ring test phase 1 have shown **that the current reliablility assessment has a high probability that different assessors would categorize** the same study differently

- During the ringtests an **insufficient experience level of participants can be excluded** as reason for the deviations in evaluations (see Fig. 1)

- This **inconsistency problem** can cause **variability in EQS derivations**

- The perception of participants is that the **current Klimisch evaluation** strongly **depends on personal expert judgement**

Preliminary conclusions of phase II (checklist approach)

- The participants evaluated the checklist system **more positively** in all categories (**accuracy, applicability, consistency and transparency**) in comparison to the Klimisch evaluation system
- **We are able to give improved guidance** first with the checklist itself, second with a more specified guidance document
- The checklist system is around **20% more strict** in study reliability evaluations
- We found a general **consistency gain for reliability**

General discussion and conclusions

- We **were able to identify the current weaknesses** in risk assessment which is mainly caused by an **inconsistency in reliability assessment**

- The newly developed checklist approach has the **potential to improve many aspects in study evaluations**

- Additionally, an **assessment of the plausibility or weight of evidence** of a study in a dataset **can lead to a higher safety in assessing critical studies** (plans for a third ringtest, 80% already indicated interest to work on this issue)

- The proposed **checklist approach is generally ready to be applied**, depending on the proposed context

Room for improvement:

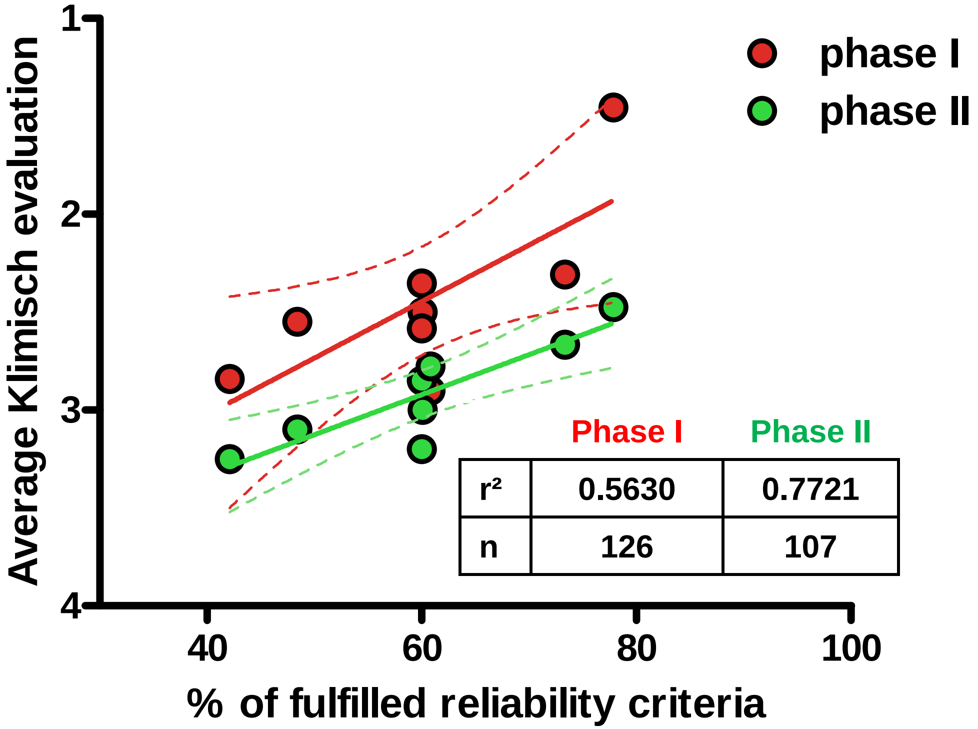
In some cases the checklist approach with the use of critical criteria **was very strict**.

One missing or not fulfilled criterion can lead to the **invalidation of a study**. High quality studies with few missing or not reported criteria might be lost, but datasets are needed for a proper risk assessment.

Discussion: Towards Klimisch 2.0?

How we can combine the advantages of the checklist evaluation system with the current need of usable data (e.g. in REACh)?

We can check and use the correlation between quality and percentage of fulfilled criteria



	Phase I	Phase II
r²	0.5630	0.7721
n	126	107

Fig. 9: Correlation analysis Klimisch code and criteria

Result:

The **mean percentage** of fulfilled criteria **correlates well with** the average **Klimisch evaluation in phase II**

Perspective:

We can **use the number of fulfilled checklist criteria** to **indicate the quality of a study**. Critical criteria could be weighted higher (e.g. x 2 times) than non-critical criteria

Such a scoring system would be **less strict** (single missed criteria will not invalidate). The **general quality level** is indicated and based on a **score for fulfilled criteria** and **available information**.

Recommendation:

Critical studies should be evaluated **with the checklist system**.

Outlook

Our intentions : - identify the current weaknesses
- improve the safety and transparency of the current assessment system
- give improved guidance
- avoiding endless discussions about single study assessment

Therefore:

- **We will publish the results in a set of publications**
- **We continue working on the plausibility and weight of evidence evaluation**
- **We need an efficient interface with the regulation to bring the advantages of the checklist approach into practice (e.g. REACh, WFD, and EMEA).**

But if we don’t try, we will loose this chance for an improvement in environmental risk assessment

Common problems need common solutions.
Thank you for your cooperation and attention !!

Outlook: Points of discussion from SETAC PAG and ERAAG 2013

In both Global Advisory Groups we found major support and agreement to put the checklist system into regulatory and scientific use.

3 main points to move forward:

- **Final revision of checklist and guidance** for advisory purposes to ECHA (Jose Tarazona, Martin Führ) and ETC (Allen Burton), (journals can use the checklists as guidance for supplementary information to allow regulatory use), expected date mid of July
- **Preparing a set of publications until next SETAC**, which will allow a **fast use for revisions of guidance documents**.
- <u>To provide you two options</u>**: Use either the more strict and consistent checklist approach and/or to use a weighting of reliability and relevance as a transparent indicator for data quality** (a score would allow to hold the current amount of data, and would provide no direct impact on AF. But a general quality gain could be driven by checklist use)

We only have to find a functional balance and timing between quality gain and strictness

Last but not least

For more information on the project visit:

http://www.oekotoxzentrum.ch/projekte/klimisch/index_EN
or
http:// www.scirap.org

Do you have any suggestions for further improvement ?

Do you have any questions ?

If yes, please do not hesitate to contact us.

contact:
Robert Kase (Robert.Kase@oekotoxzentrum.ch)
Muris Korkaric (Muris.Korkaric@eawag.ch)
Marlene Agerstrand (Marlene.Agerstrand@itm.su.se)
Caroline Moermond (Caroline.Moermond@rivm.nl)

Source: Gerd Maack, SETAC 2010 Sevilla, Frosch_Roman_Arcea_Fotolia_1260576_Subscription_L.jpg