APPENDIX 6

Comparison of risk assessment between experts and ASSYST_NTM

Vergleich der Risikobewertungen zwischen Experten und ASSYST_NTM
(auszugsweise entnommen aus / 1 /)

Gesamtüberblick über die Abweichungen

Die durchschnittliche Abweichung (Mittelwerte der Beträge der Differenzen) der Bewertungen der Experten von den ASSYST_NTM(vormals NARIDAS) -Werten liegt über alle Szenarien hinweg bei 0,20.

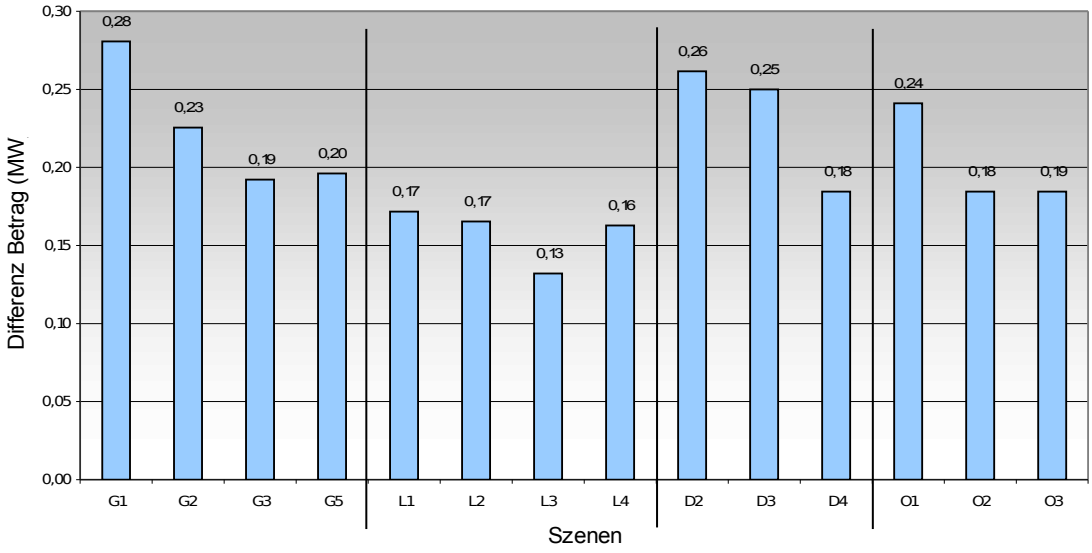

Abb. 1: Szenenvergleich der Beträge der mittleren Abweichungen zwischen VPn und ASSYST_NTM über alle Teilrisiken.

Tendenziell sinken die Abweichungen im zeitlichen Verlauf innerhalb der einzelnen Szenarien, was für einen gewissen Lerneffekt bzw. eine Annäherung der Bewertungen der Experten an ASSYST_NTM spricht (Abb. 1).
Abb. 2 zeigt die mittleren Abweichungen zwischen Experten und dem ASSYST_NTM über alle Szenen für die einzelnen Teilrisiken. Am geringsten weichen die Bewertungen für das Teilrisiko MANNING, am meisten für COLLISION und ECONOMY ab. Das Teilrisiko ECONOMY wird von den Experten in 12 von 14 Szenen geringer eingeschätzt als durch das ASSYST_NTM. Die Ursachen dieser Abweichungen werden nachfolgend erörtert, indem für jedes der acht Teilrisiken die Bewertungen von Experten und ASSYST_NTM für die einzelnen Szenen verglichen und die Begründungen für die Differenzen aus den verbalen Protokollen zusammenfassend dargestellt werden.

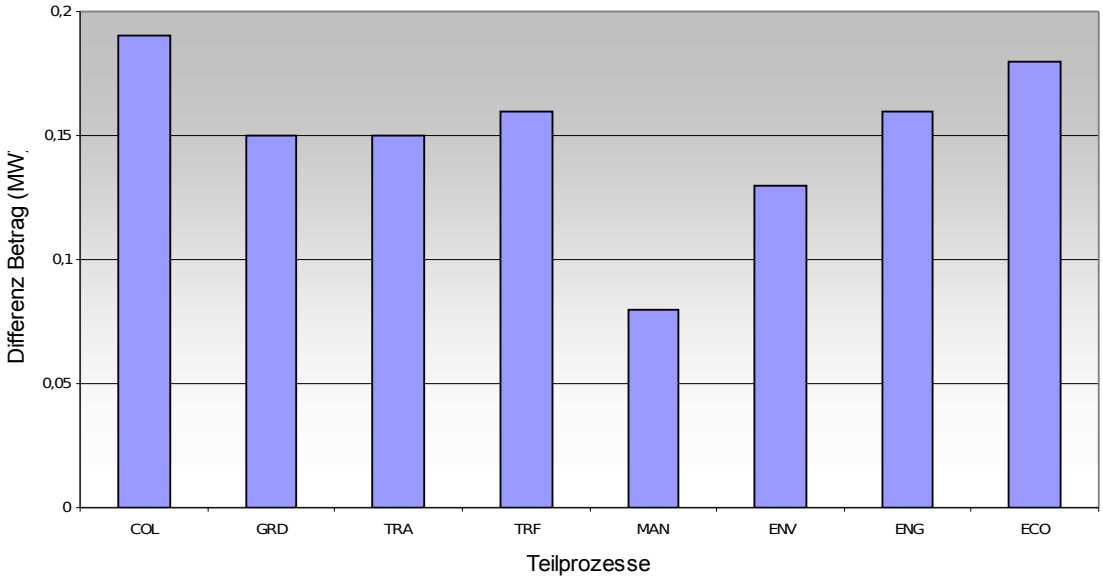

Abb. 2: Vergleich der Abweichungen der Mittelwerte zwischen VPn und ASSYST_NTM über alle Szenen nach Teilrisiken.

Einzelanalyse der Teilrisiken

Insgesamt schätzen die Experten die Teilrisiken im Mittel geringer ein als das ASSYST_NTM, so dass eine grundsätzliche Tendenz des Systems zur Überschätzung der Risiken zu bestehen scheint (s.u.).

COLLISION

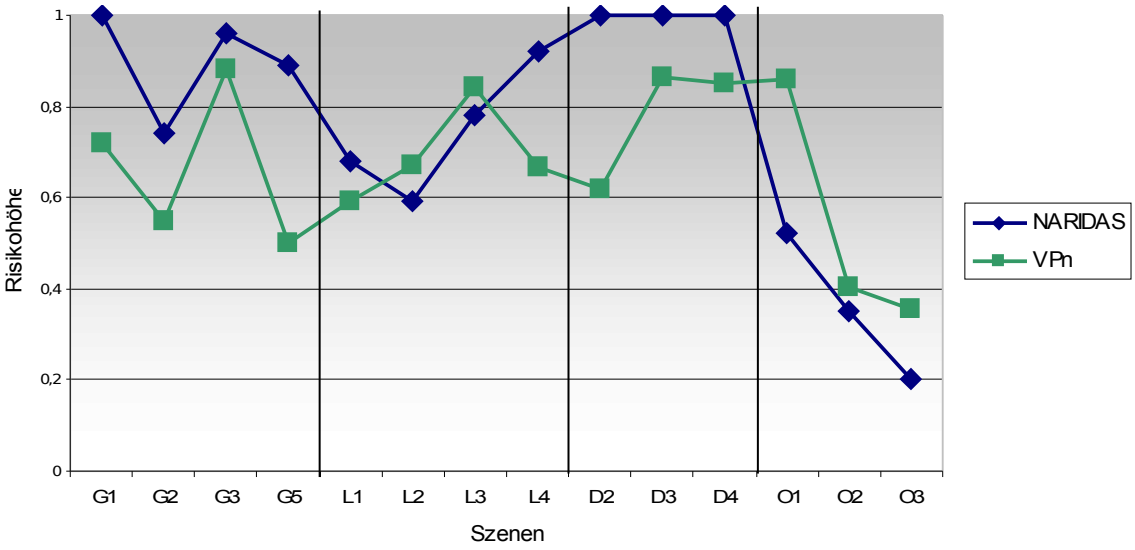

Abb.3: Vergleich der Bewertungen des Teilrisikos COLLISION von Experten (VPn) und ASSYST_NTM

Zusammenfassende Beurteilung von COLLISION: Inhaltliche Bedeutung und Berechnungsgrundlagen dieses Teilprozesses werden von den Experten schnell erfasst. Allerdings bereitet die statische Bewertung dieser sehr dynamischen Größe öfters Schwierigkeiten. Relativ häufig tendieren die Experten dazu, das COLLISION Risiko etwas geringer zu bewerten als ASSYST_NTM. Daher sollte überprüft werden, ob z.B. die „cpa good seamanship“ in bestimmten *Navigation Modes* heruntergesetzt werden könnte.

GROUNDING

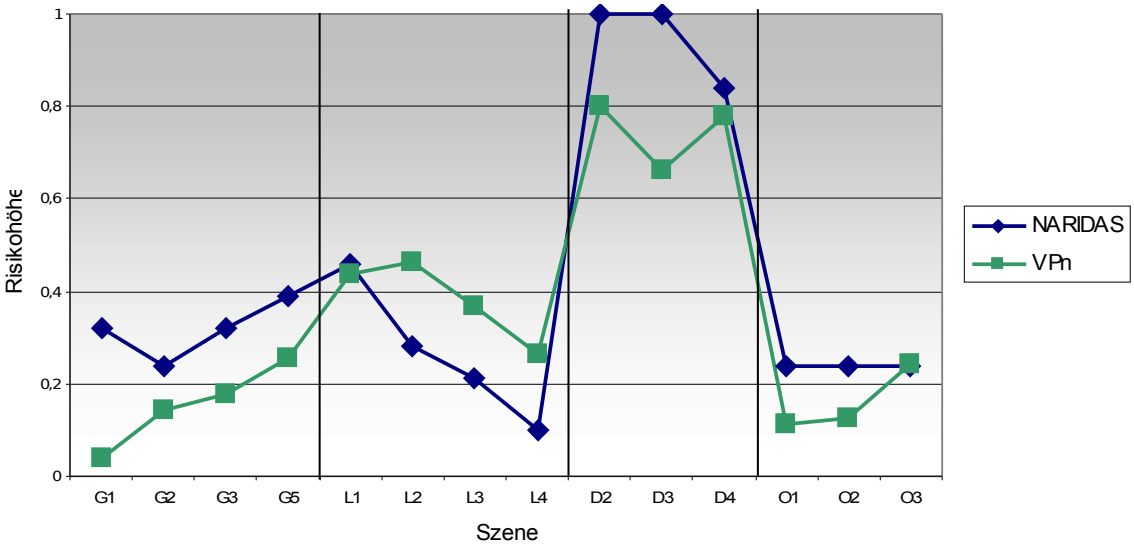

Abb.4: Vergleich der Bewertungen des Teilrisikos GROUNDING von Experten (VPn) und ASSYST_NTM

Zusammenfassende Beurteilung von GROUNDING: Die Experten haben mit dem Verständnis der inhaltlichen Bedeutung und Berechnungsgrundlagen dieses Teilprozesses keine Schwierigkeiten, die Abweichungen zwischen Experten und ASSYST_NTM sind relativ gering. Es bleibt zu berücksichtigen, dass dieser Parameter nur in Situationen mit einer möglichen Grundberührung relevant ist, d.h. z.B. im Navigation Mode „Open Sea“ keine Rolle spielt.

TRACK

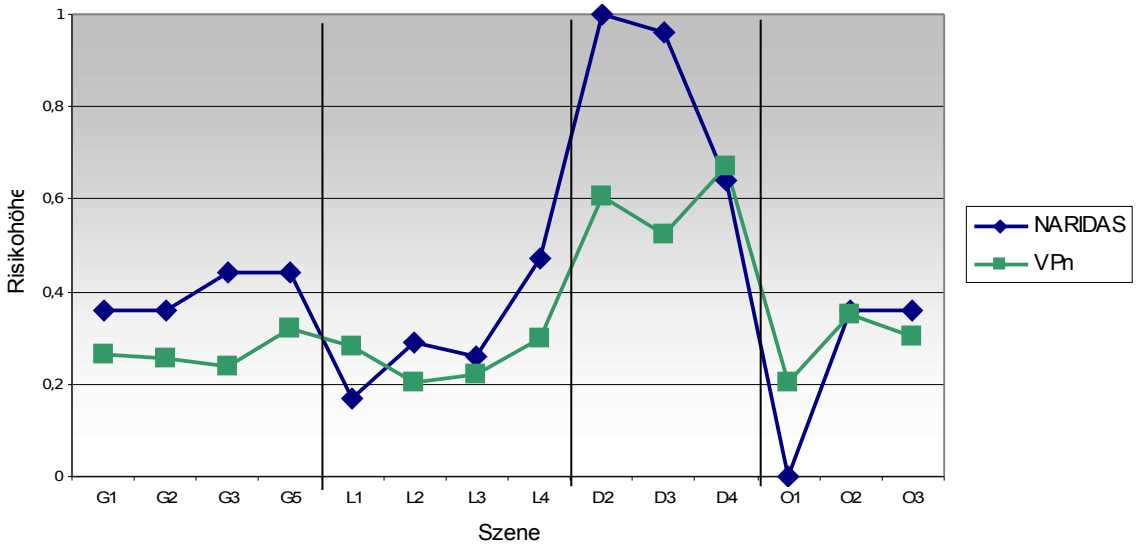

Abb.5: Vergleich der Bewertungen des Teilrisikos TRACK von Experten (VPn) und ASSYST_NTM.

Zusammenfassende Beurteilung von TRACK: Die genaue Bedeutung und Berechnung dieses Teilrisikos ist nicht allen Experten auf Anhieb verständlich. Aufgrund ihres gewohnten Sprachgebrauchs wird dieses Risiko von einigen Experten ausschließlich auf die Bahnabweichung bezogen. Allerdings erscheint den Experten die „intelligentere“ Berechnung dieses Teilrisikos durch ASSYST_NTM in der Diskussion einleuchtend.

TRAFFIC

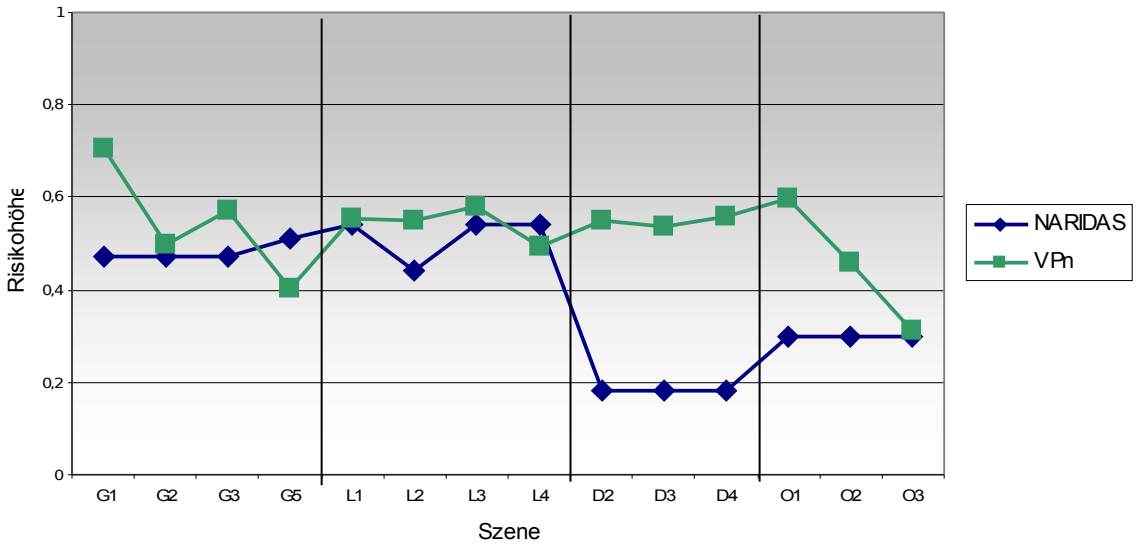

Abb.6: Vergleich der Bewertungen des Teilrisikos TRAFFIC von Experten (VPn) und ASSYST_NTM.

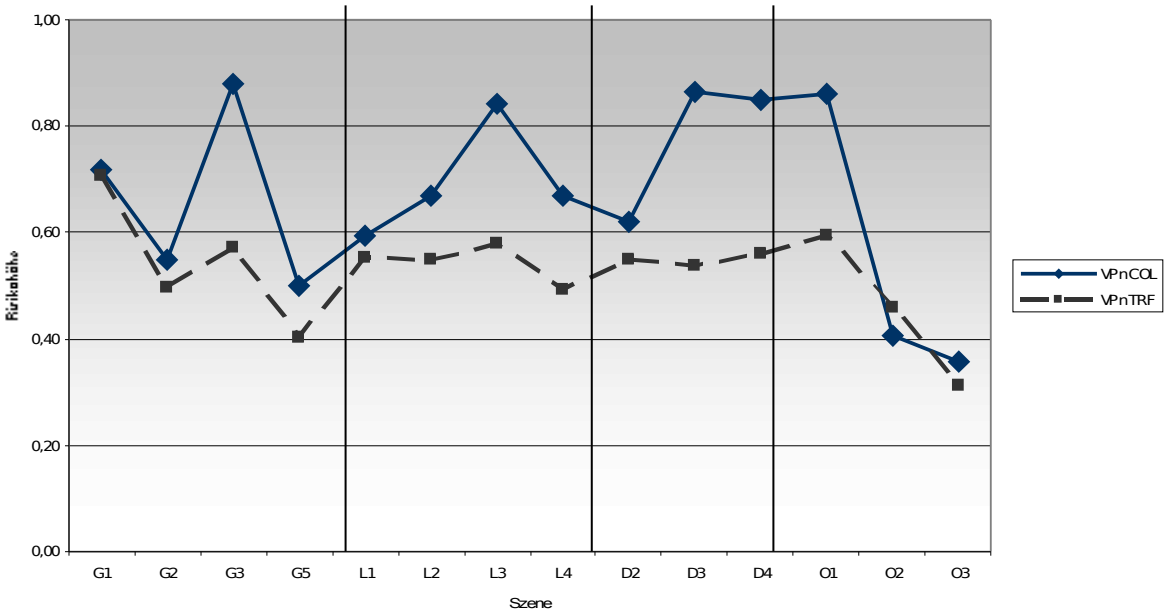

Abb.6a: Vergleich der Expertenbewertungen für COLLISION und TRAFFIC

Zusammenfassende Beurteilung von TRAFFIC: Die Abgrenzung dieses Teilrisikos zu COLLISION erscheint einigen Experten unklar. Der zusätzliche Nutzen der Anzeige von TRAFFIC zur Unterstützung der Schiffsführung wird aus diesem Grund teilweise angezweifelt.

MANNING

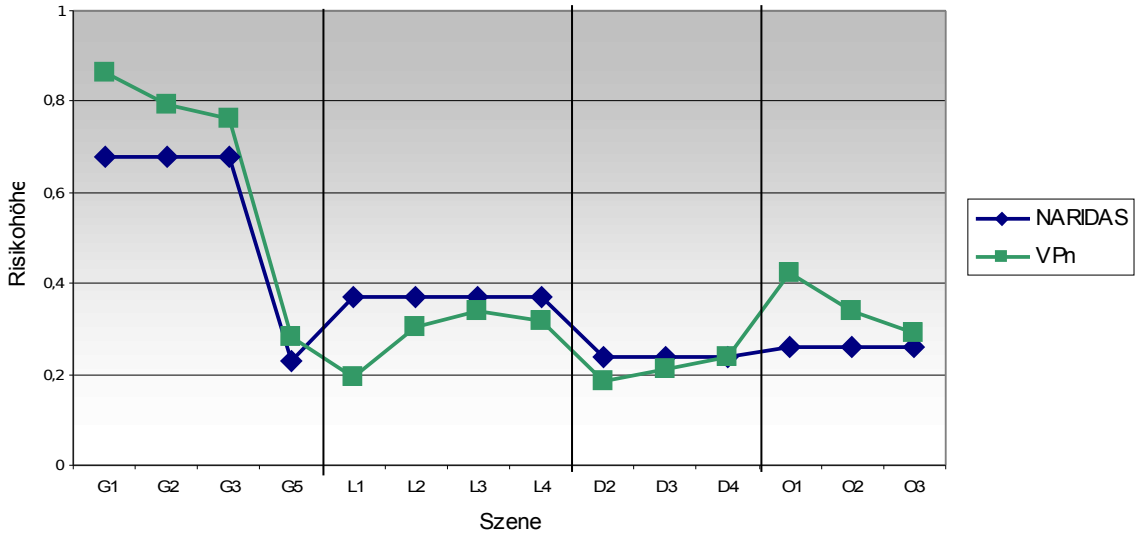

Abb.7: Vergleich der Bewertungen des Teilrisikos MANNING von Experten (VPn) und ASSYST_NTM.

Zusammenfassende Beurteilung von MANNING: Bedeutung und Berechnung dieses Teilrisikos sind für die Experten gut nachvollziehbar, die Übereinstimmung mit ASSYST_NTM ist relativ hoch. Das Unterstützungspotenzial der Anzeige von MANNING erscheint jedoch umstritten, da Zusammensetzung und Zustand der Brückenbesatzung zumeist als gegebene Rahmenbedingungen betrachtet werden, und in der operativen Schiffsführung nur ein sehr geringer Handlungsspielraum zur Verringerung dieses Teilrisikos gesehen wird.

ENVIRONMENT

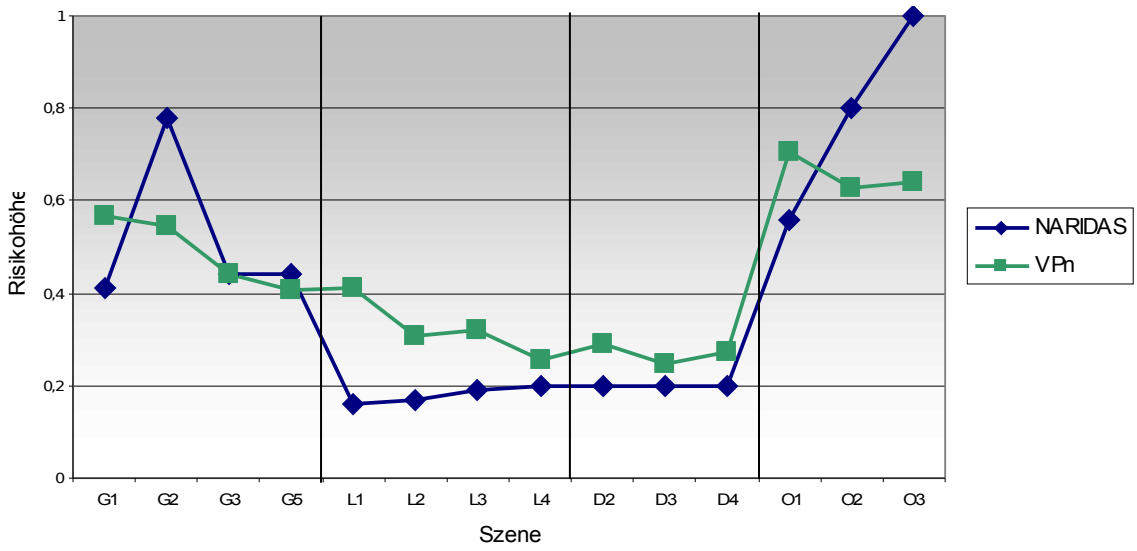

Abb.8: Vergleich der Bewertungen des Teilrisikos ENVIRONMENT von Experten (VPn) und ASSYST_NTM

Zusammenfassende Beurteilung von ENVIRONMENT: Inhaltliche Bedeutung und Berechnungsgrundlagen dieses Teilrisikos erscheinen den Experten gut verständlich, die Übereinstimmung mit ASSYST_NTM ist relativ hoch. Die Berechnung eines Risikowertes für die Resonanz wird von den meisten Experten ausdrücklich begrüßt.

ENGINE

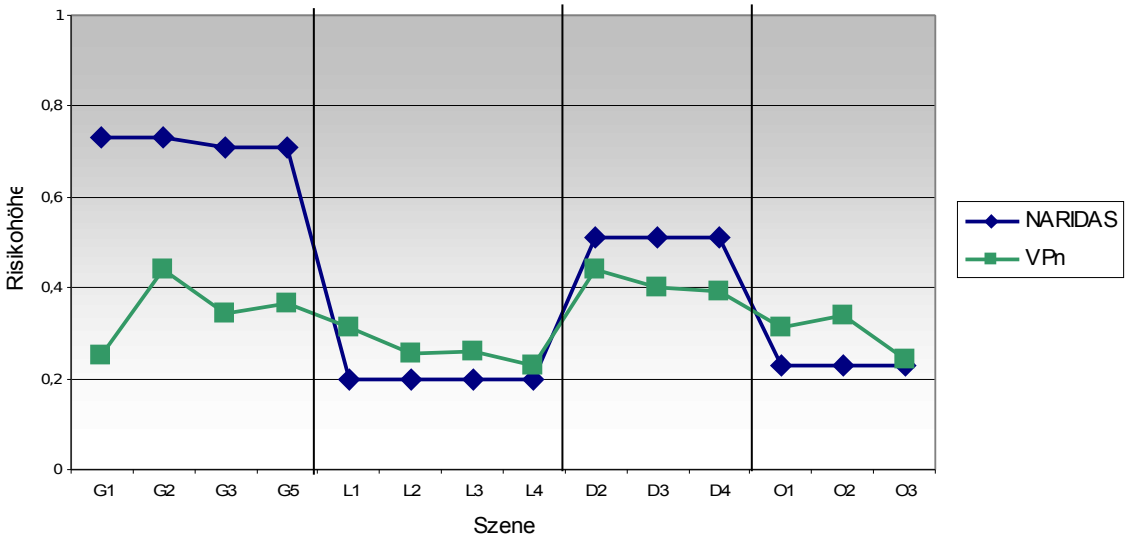

Abb.9: Vergleich der Bewertungen des Teilrisikos ENGINE von Experten (VPn) und ASSYST_NTM

Zusammenfassende Bewertung von ENGINE: Da sich der Algorithmus noch in der Entwicklung befindet, kann aufgrund der vorliegenden Untersuchung über diesen Teilprozess keine Aussage getroffen werden.

ECONOMY

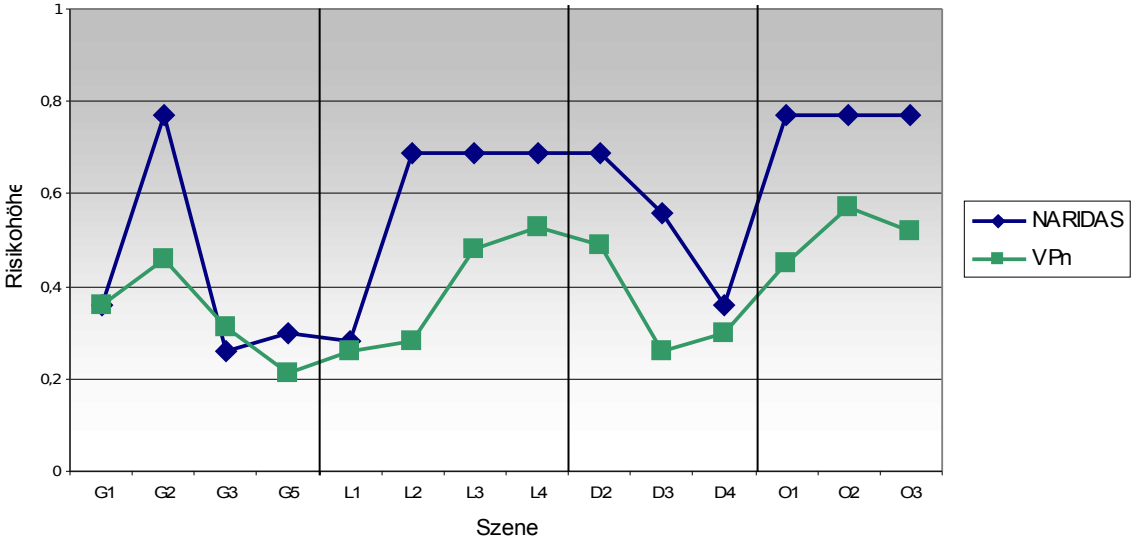

Abb.10: Vergleich der Bewertungen des Teilrisikos ECONOMY von Experten (VPn) und ASSYST_NTM

Eine Aufzeichnung der Ökonomiewerte wird von einigen Experten generell mit Skepsis betrachtet, da dadurch eine zusätzliche Kontrolle durch Reeder und Vorgesetzte erwartet wird.

Zusammenfassende Bewertung von ECONOMY: Die Konzeption dieses Teilrisikos ist für die Experten prinzipiell nachvollziehbar. Allerdings erscheint die Nützlichkeit bzw. das Unterstützungspotenzial der Anzeige dieses Wertes für die operative Schiffsführung sehr umstritten und stößt auf hohe Skepsis.

Analyse von Spezifität und Sensitivität der Risikoberechnungen von ASSYST_NTM

Zwei wichtige Kriterien für die Evaluation der in ASSYST_NTM implementierten Algorithmen zur Risikoberechnung sind **Spezifität** und **Sensitivität**. Spezifität wird hier verstanden als das Ausmaß, in dem vom System erkannte kritische Zustände auch von den Experten als kritisch beurteilt werden. Hat ein System eine niedrige Spezifität, bedeutet dies, dass es eine hohe Rate von Fehlalarmen produziert. Im Gegensatz dazu kann Sensitivität definiert werden als das Ausmaß, in dem das System von den Experten identifizierte kritische Zustände ebenfalls als kritisch bewertet. Eine niedrige Sensitivität bedeutet demzufolge, dass das System nicht in der Lage ist, vor gefährlichen Situationen zu warnen. Während eine niedrige Spezifität in erster Linie eine schlechte Akzeptanz des Systems erwarten lässt (da es aus Sicht der Benutzer ohne Anlass warnt), ist eine niedrige Sensitivität sowohl hinsichtlich der Akzeptanz als auch hinsichtlich der Zuverlässigkeit eines Systems zur Risikobewertung verheerend, da ein wesentlicher Zweck des Systems, die Warnung vor Gefahren, verfehlt wird.
Für die nachfolgende Analyse wird Spezifität über den Anteil an Fehlalarmen erfasst. Ein Fehlalarm (*false alarm*) wird operationalisiert als ein Fall, in dem mehr als der Hälfte der Experten das Teilrisiko um mehr als 0,2 Punkte geringer bewertet als ASSYST_NTM. Sensitivität wird über den Anteil an Auslassungen erfasst. Eine Auslassung (*miss*) wird operationalisiert als ein Fall, in dem mehr als die Hälfte der Experten ein Teilrisiko um mehr als 0,2 Punkte höher bewertet als ASSYST_NTM.
Nach dieser Definition liegt die *false alarm* Quote in der vorliegenden Untersuchung bei 17%, die *miss* Quote bei 7,1%. Allerdings sind hierbei alle Fälle auf der gesamten Risikoskala berücksichtigt. Da ein „Fehlalarm“ in einem niedrigen Risikobereich (z.B. „no risk“, „low risk“) als relativ unkritisch betrachtet werden kann, sollten jedoch für eine derartige Analyse in erster Linie die kritischen Zustände betrachtet werden. Wertet man lediglich ASSYST_NTM Risikowerte >0,8 als kritisch, so ergibt sich eine *false alarm* Quote von 7,1% (bzw. eine Spezifität von 92,9%). Betrachtet man dementsprechend lediglich diejenigen Auslassungen als kritisch, in denen über die Hälfte der Experten das Risiko >0,8 einschätzt, liegt die *miss* Quote bei 0,9% (bzw. die Sensitivität bei 99,1%).
In den Tab.3 und 4 sind für alle Teilrisiken diejenigen Szenen aufgelistet, in denen ein *false alarm* bzw. *miss* auftrat, sowohl für die Pilotstudie (NHWS) als auch für die aktuelle Untersuchung (ASSYST_NTM). In Tab.3 sind diejenigen Szenen fett gedruckt, die als „kritische“ *misses* zu werten sind, d.h. in denen mehr als die Hälfte der Experten das Risiko >0,8 einschätzt (in Klammern hinter der Szenenbezeichnung steht die Anzahl der Experten, die das Risiko >0,8 einschätzt; steht keine Zahl in Klammern, ist dies für keinen Experten der Fall). In Tab.4 sind diejenigen Szenen fett gedruckt, die als „kritische“ *false alarms* zu werten sind, d.h. in denen ASSYST_NTM das Risiko >0,8 bewertet, während über die Hälfte der Experten um mehr als 0,2 Punkte unter ASSYST_NTM liegt.

	NHWS	ASSYST_NTM
COL/TAR	L2(2), **O1(7)**	**O1 (5)**
GRD/SPD		
TRA	D1(2)	
TRF	G1(2), D1(1), D2, D3, D4	G1(3), D2(1), D3, D4(1), O1(1)
MAN/HUM	D1(3), O1(1)	
ENV		L1, O1(3)
ENG/AVAIL		
ECO		

Tab.3: Vergleich der Szenen, in denen *misses* auftreten NHWS (Studie 2004, N=9) vs. ASSYST_NTM (Studie 2005, N=7)

Insgesamt zeigt sich in Tab. 3 für beide Untersuchungen lediglich ein einziger kritischer *miss*, der sich auf denselben Fall, nämlich die Beurteilung von COLLISION in der Szene *Open Sea 1* bezieht. Die Analyse der verbalen Protokolle zeigt, dass die höhere Bewertung dieses Teilrisikos in dieser Situation darin begründet liegt, dass die Experten das Risiko stärker vorausschauend bewerten, auf offener See einen größeren Passierabstand bevorzugen, und auch die relativ schlechten Umweltbedingungen berücksichtigen.
Auffällig ist zudem die relativ hohe (unkritisch) *miss* Rate für das Teilrisiko TRAFFIC, die sowohl in der ersten Studie als auch in der Folgeuntersuchung 2005 auftritt. Hier spiegelt sich u.a. die oben angesprochene Problematik wider, dass die Experten TRAFFIC in Bezug auf das Gefährdungspotential des Verkehrs und damit in enger Verbindung mit dem Teilrisiko COLLISION bewerten.

	NHWS	ASSYST_NTM
COL/TAR	**G5**, **D2**, O2	G2, **G5**, **L4**, **D2**
GRD/SPD		G1, **D3**
TRA	L2, L4, L5	G3, **D2**, **D3**
TRF		
MAN/HUM		
ENV		**O2**, **O3**
ENG/AVAIL	G5, D1	G1, G2, G3, G5
ECO	L2, L5, D2	L2, D3, O1, O3

Tab.4: Vergleich der Szenen, in denen *false alarms* auftreten NHWS (Studie 2004) vs. ASSYST_NTM (Studie 2005).

Die Anzahl der Szenen, in denen *false alarms* auftreten, hat sich in der Studie 2005 deutlich erhöht. Da an den beiden Studien bis auf eine Ausnahme unterschiedliche Experten beteiligt waren, lässt sich nicht klären, ob diese Zunahme auf Eigenschaften der Stichprobe oder auf die (allerdings geringfügigen) Änderungen der ASSYST_NTM-Algorithmen zurückzuführen ist. Ein möglicher Stichproben-Effekt könnte darin begründet liegen, dass sich die an der ersten Untersuchung teilnehmenden Studenten eher an die Bewertungen des Systems angepasst haben als die erfahreneren (und dadurch vielleicht eher „stureren“) Nautiker in der aktuellen Untersuchung.
Die meisten kritischen *false alarms* treten für das Teilrisiko COLLISION (2005) bzw. TARGET (2004) auf, in beiden Untersuchungen in den Szenen *Gibraltar 5* und *Dover 2*. Dafür kann als Begründung auf die o.g. Auswertung der Verbaldaten zurückgegriffen werden. In der Szene *Livorno 4* gibt das NHWS für TARGET einen Wert von 0,76, das

ASSYST_NTM für COLLISION den Wert 0,92 an. Betrachtet man die Bewertungen der einzelnen Experten der Studie 2005 liegen diese eher um den Wert von 0,76, d.h. sie hätten mit den „alten“ Werten keinen *false alarm* produziert. Eventuell sollten hier die inzwischen vorgenommenen Veränderungen in der Wissensbasis noch einmal überdacht werden.
Für den Parameter GROUNDING/SPEED ist die Bewertung des Systems in den Szenen *Gibraltar 1* und *Dover 3* gleich geblieben, d.h. die Experten der aktuellen Studie beurteilen dieses Risiko in diesen Szenen niedriger als ihre Kollegen in der vorigen Untersuchung.
Positiv hat sich die Modifikation der Risikowerte für TRACK im Szenario *Livorno* ausgewirkt. Hier treten 2005 keine *false alarms* mehr auf. Für die Abweichungen in den Szenen *Dover 2* und *3*, in denen das System unverändert einen Risikowert von 1 angibt, finden sich die Begründungen in der Analyse des Risikos TRACK (s.o.). Auch die „neuen“ false alarms für ENVIRNOMENT im Szenario Open Sea sind auf unterschiedliche Bewertungen der Experten in den beiden Untersuchungen zurückzuführen.
Insgesamt muss betont werden, dass sowohl den Fragebogendaten als auch den Audioprotokollen zu entnehmen ist, dass die Experten wiederholt Schwierigkeiten mit der Einteilung der Risikoskala hatten. Ein gute Ansatzpunkt ist es sicherlich, die vom System verwendete Skala zumindest im unteren Bereich (bis 0,4) zu stauchen. Grundsätzlich sollte weiterhin über eine handlungsleitende Kategorisierung der Risikowerte nachgedacht werden, um das Informationsangebot durch ASSYST_NTM zu reduzieren. Aus ergonomischer Sicht erscheint z.B. eine Reduktion der qualitativen Risikokategorien auf drei (z.B. „low risk“ – „caution“ – „danger“) hinreichend und „kognitiv ökonomisch“.
Zu berücksichtigen sind bei der Analyse der Risikowerte auch die wenig standardisierten Untersuchungsbedingungen in beiden Studien. So können sich unterschiedliche Hilfestellungen durch die Versuchsleiter bei der Risikobewertung beispielsweise auf die Genauigkeit der Einschätzungen auswirken. Beispielsweise kann in den Szenen *Open Sea 1* und *2* ein Übergehen der Daten, die eine Resonanzschwingung anzeigen, sich in einem *false alarm* durch das System widerspiegeln (der in diesem Falle allerdings kein Fehlalarm sondern eine korrekte Warnung wäre), während Experten, die auf diese Werte von den Versuchsleitern konkret hingewiesen werden, das ENVIRONMENT Risiko ähnlich hoch wie das System einschätzen.

<u>Zusammenfassende Bewertung der Spezifität und Sensitivität</u>: Die Sensitivität von ASSYST_NTM liegt in den beurteilten Situationen bei nahezu 100%, die Spezifität erreicht einen Wert von 92,9%. Diese Werte sprechen für eine hohe Zuverlässigkeit der Algorithmen. Für die Optimierung des Systems ist eine weitere Verringerung des Anteils an Fehlalarmen bei einer gleichbleibend hohen Sensitivität anzustreben.

Validität der Risikoberechnungen

Über alle Szenen und Teilprozesse ergibt sich für die sieben Experten und ASSYST_NTM ein Cronbachs Alpha von 0,89, was einer sehr hohen Übereinstimmung entspricht.
Tab. 5 zeigt die bivariaten Korrelationskoeffizienten (Spearman Rho) zwischen den einzelnen Experten und ASSYST_NTM. Für jeden einzelnen Experten ergibt sich über alle Szenen und Teilprozesse eine signifikante Korrelation mit ASSYST_NTM ($p<.01$). Allerdings ergeben sich für zwei Experten (VP 1 und 7) deutlich niedrigere Korrelationskoeffizienten als für die übrigen, was darauf hinweist, dass die Experten unterschiedlich stark mit den

Risikoberechnungen des Systems übereinstimmen. Da sich jedoch kein Zusammenhang zwischen der Höhe der Übereinstimmung mit ASSYST_NTM und der Bewertung der Gebrauchstauglichkeit (SUS Score) ergibt, erscheint die Übereinstimmung mit den Risikobewertungen die Gesamtbeurteilung des Systems nicht zu beeinflussen, so dass diese Unterschiede vernachlässigt werden können. Somit zeigt sich, dass die Risikoberechnungen von ASSYST_NTM der Situationseinschätzung von Experten insgesamt gut entsprechen und nachvollziehbar sind.

	VP 1	VP 2	VP 3	VP 4	VP 5	VP 6	VP 7	MW VPn
Korrelation mit ASSYST_ NTM	.27	.63	.50	.56	.67	.55	.31	.66

Tab. 5: Bivariate Korrelationen (Spearman Rho) zwischen Experten (VPn) und ASSYST_NTM

Ein Vergleich zeigt, dass die Übereinstimmung der Experten mit ASSYST_NTM in der vorliegenden Untersuchung etwas niedriger ist als in der vorangegangenen Pilotstudie. Cronbachs Alpha sinkt von 0,94 auf 0,89, der Mittelwert der Beträge der absoluten Abweichungen zwischen Experten und ASSYST_NTM steigt von 0,17 auf 0,20. Dieser Anstieg findet sich bei nahezu allen Teilrisiken (Abb. 11) und Szenen (Abb. 12).

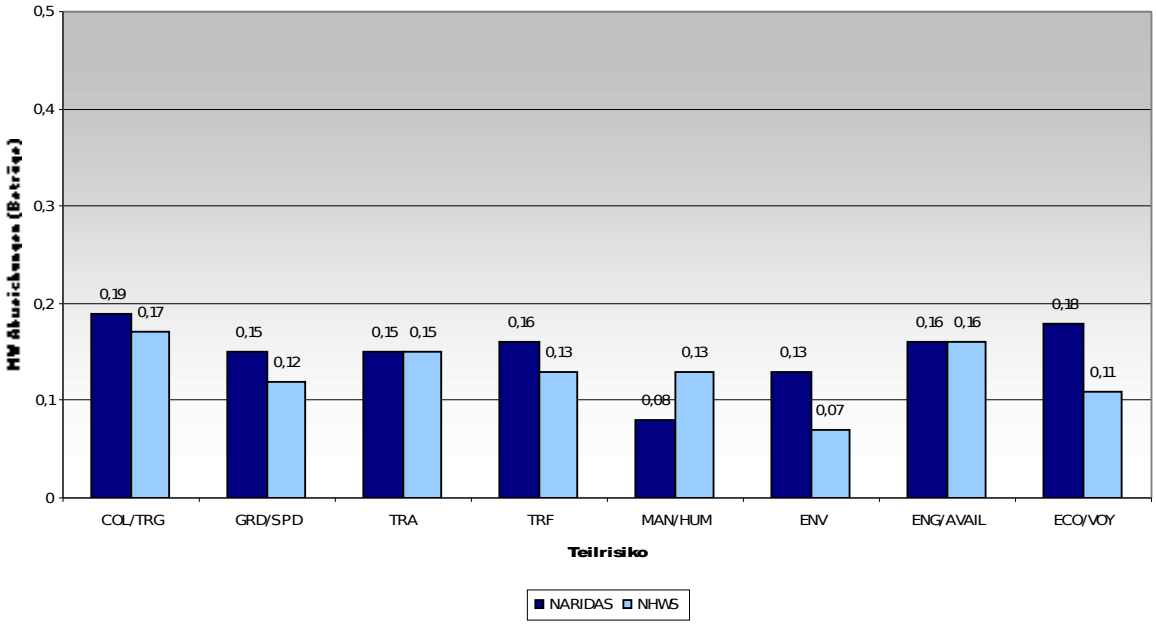

Abb. 11 Vergleich der Mittelwerte der Beträge der absoluten Abweichungen zwischen aktueller Untersuchung (ASSYST_NTM) und Pilotstudie (NHWS) in den Teilprozessen

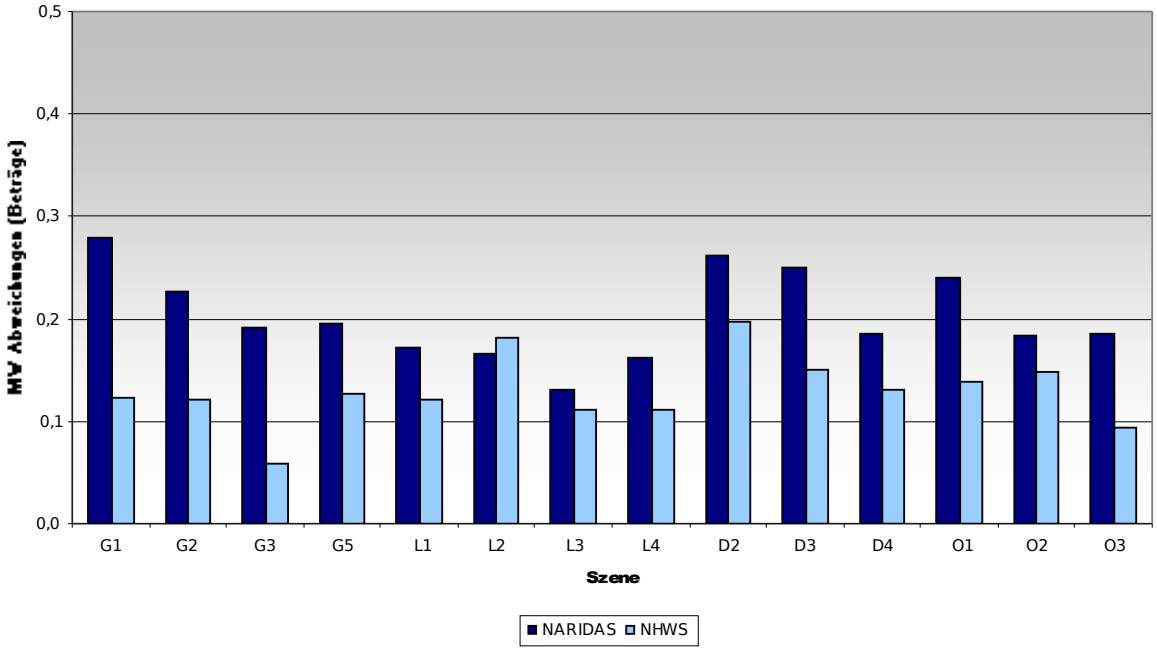

Abb. 12 Vergleich der Mittelwerte der Beträge der absoluten Abweichungen zwischen aktueller Untersuchung (ASSYST_NTM) und Pilotstudie (NHWS) in den Szenen

Tatsächlich liegt Cronbachs Alpha in beiden Untersuchungen in einem hohen Bereich (>0.8), und die bivariaten Korrelationen zwischen Experten und System sind durchweg signifikant (p<.01).

LITERATUR :

/ 1 / Gauss, B. : Evaluation von Situational Risk Assessment Systemen
Entwicklung eines Rahmenkonzepts und Demonstration seiner Anwendbarkeit im Bereich der Schiffsführung
Dissertation – 14. März 2008
Fakultät V – Verkehrs- und Maschinensysteme der Technischen Universität Berlin

/ 2 / Kersandt, D. : Qualitätsbestimmung von Schiffsführungsprozessen (Hauptbericht)
- QUASNAV -
F/E – Bericht, Rostock, Juni 2010