Review of Marine Benthic Indicators and Metrics for the WFD and design of an optimized BEQI

Deltares

Review of Marine Benthic Indicators and Metrics for the WFD and design of an optimized BEQI

dr. A.R. Boon
dr. A. Gittenberger (Gimaris)
dr. W.M.G.M. van Loon (RWS Waterdienst)

1203801-000

© Deltares, 2011

Deltares

Title
Review of Marine Benthic Indicators and Metrics for the WFD and design of an optimized BEQI

Project	**Reference**	**Pages**
1203801-000	1203801-000-ZKS-0006	59

Keywords
Benthic indicators, Water Framework Directive, AMBI, BEQI, NQI, DKI, BAT, ITI, Species richness, Shannon-Wiener.

Summary
In this report, indicators and metrics that have been developed and applied within the context of the Water Framework Directive have been reviewed. The goal of this review was to prepare the optimisation of the Dutch indicator for benthic quality. Indicators and metrics were evaluated by set-up, calibration and performance criteria and scored numerically by the authors of this report. This resulted in a ranking of indicators and metrics in each of the three WFD categories, with a highest (top 2) ranking for Shannon-Wiener and Margaleff's *d* in the category "diversity indicator", total density and abundance-biomass ratio in the category "abundance indicator" and AMBI and ITI in the category "sensitive/opportunistic species". The two best scoring (multi)metrics in this evaluation were DKI and the NQI, which make use of the best scoring indicators mentioned above. Suggestions are made for the composition of a new Dutch benthic quality metric, based on the best scoring indicators and metrics.

Version	Date	Author	Initials	Review	Initials	Approval	Initials
1	Jun 2010	dr. A.R. Boon		dr. T.C. Prins		T. Schilperoort	b/a
2	Sep. 2010	dr. W.M.G.M. van Loon					
3	Nov. 2010	dr. A. Gittenberger					
4	Jan. 2011	dr. W.M.G M. van Loon					
5	Feb. 2011	dr. A.R. Boon					b/a

State
final

Contents

List of abbreviations

A	Abundance, expressed as numbers or densities.
ADF	Alpha Diversity Fisher.
AeTI	Aestuar Type Indicator.
AeTV	(German) Aestuar Type Verfahren.
AMBI	(Spanish) Aztec Marine Biotic Index.
BAT	(Portuguese) Benthic Assessment Tool.
BEQI	(Dutch and Belgian) Benthic Ecosystem Quality Index.
BOPA	Benthic Opportunistic Polychaetes Amphipods index.
BQI	(Swedish) Benthic Quality Index.
Bray Curtis index	Similarity index which compares communities based on species and their relative densities.
DKI	Danish Quality Index.
ES50	Expected number of species at the count of 50 individuals.
H’	Shannon Wiener Index. This index is also referred to as the Shannon index.
IQI	(British) Infaunal Quality Index.
ITI	Infaunal trophic index. Index based on the classification of species in four trophic groups.
λ/Lambda	Simpson index.
Marbit	(German) Marine Biotic Index Tool
m-AMBI	Multivariate AMBI.
NQI	Norwegian Quality Index.
r/K-strategist	Classification of species based on their survival strategy.
S	Species richness, the number of species.
SN	Diversity index, defined as ln(S)/ln(ln(N))

1 Introduction

1.1 Problem definitions

In the context of the EU Water Framework Directive, most EU countries have developed a metric, here defined as a combination of a few indicators/indices, to model and estimate the ecological status of the marine benthic ecosystem. Such WFD indicators have also been developed for more limited ecological groups such as phytoplankton and fish. The index for the benthic ecosystem is one of the most complex indices, covering a whole subsystem instead of a species group or single process. Therefore a benthic index is not easy to set up. When reviewing the results from different countries, in and outside the EU, so far a large list of indicators and indices has emerged (Josefson et al. 2009, Dauvin et al. 2010), with each index having its own characteristics. Interestingly, a comparable benthic indicator is currently being sought for the environmental status descriptor "Seafloor integrity" in the EU Marine Strategy Framework Directive (MSFD)[1].

The Waterdienst of Rijkswaterstaat is responsible for the development and intercalibration of the Dutch method to assess the marine benthos state for the WFD. The current Dutch method is the BEQI, the Benthic Quality Index. The BEQI was developed by the NIOO (Van Hoeij et al. 2007) and was translated into the formal Dutch WFD metric (Van der Molen & Pot 2006, Twisk et al. 2009).

A schematic overview of the BEQI is given in figure 1.1 (from Twisk et al. 2009). It can be seen in this figure that the BEQI assessment consists of three levels. Level 1 is an ecosystem level, level 2 a global habitat level (may contain several ecotopes per global habitat) and level 3 the community level. The intercalibration process is focused on level 3, since the other European countries do not assess level 1 and 2.

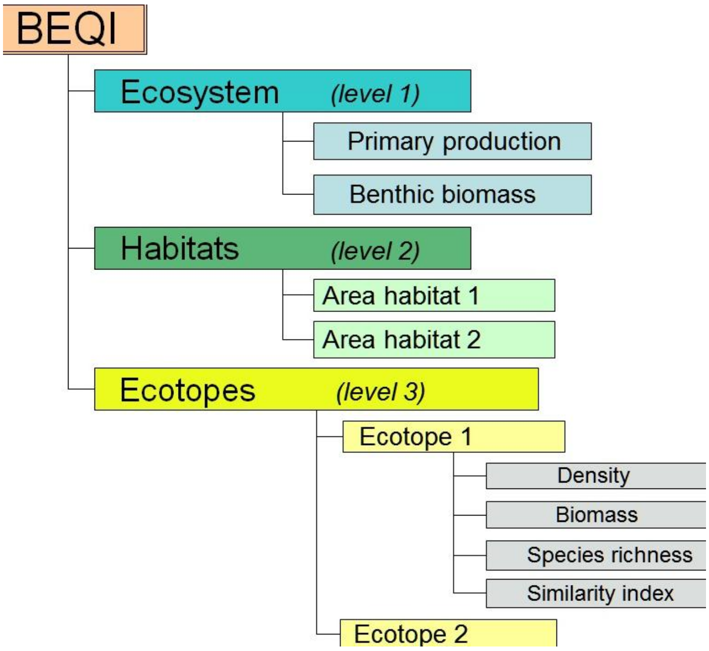


Figure 1.1: schematic overview of the BEQI.

[1] *MSFD text on http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:164:0019:0040:EN:PDF*

1.2 Review questions

In the past three years several questions have been raised by the RWS Waterdienst on the ecological validity, calibration and intercalibration of the BEQI. Examples of these questions are:

- Is the level 1 assessment (ratio primary production/total benthic biomass) necessary and can it be calculated reliably?
- Can the indicator total density be calibrated to human pressures?
- Can the indicator total biomass be calibrated to human pressures?
- Which diversity index is most suited for the BEQI?
- Is the Bray Curtis Similarity Index valid as an indicator for sensitive/opportunistic species? If not, which indicators are suited?
- Which calibration method is most suitable for the BEQI?

The main question here is: what is an optimal[2] assessment method for the marine invertebrate benthos in transitional and coastal waters for the EU Water Framework Directive (also having potential for application in the EU Marine Framework Strategy Directive and large infrastructural projects from Rijkswaterstaat).

1.3 Review objectives

In order to get a broad overview of available benthic indicators and indices before answering the questions above, this report gives a synopsis of the most important of these indices, focusing on those currently being used within the EU and which have been used in the Netherlands recently. It looks at the methodology of the indicators, their calibration and validation, and applications in European indices or multi-metrics so far. Based on this review, an improved and simplified setup of the BEQI is proposed. It is obvious that, when setting up a new method for such an index, additional work needs to be done for calibrating the index. This is outside the scope of this report however and is planned as a follow-up project in the beginning of 2011.
First, in chapter 2, a concise overview is given of benthic ecosystem characteristics in transitional and coastal waters, and the human pressures and impacts on the benthic ecosystem in the Dutch situation. Next, in chapter 3, a short background description is given of the definitions for benthic ecosystem status in the WFD, and the selection of relevant indices in this context. In chapter 4, both a description and evaluation is given of the most relevant indices and metrics that have been developed so far. The indices are evaluated using a set of evaluation criteria. In the last chapter, the evaluation is summarised, discussed and an optimized BEQI method is proposed.

In the last decade, a large number of publications have appeared on the WFD assessment of marine macrozoobenthos. It needs to be stressed that due to time constraints, the evaluation could not encompass all literature on this subject, but most recent WFD literature has been covered. Furthermore, this review does not give an absolute qualification of the indicators. The evaluation is based on a limited review of literature and an expert judgement by the authors and is meant for an intercomparison. As such, the results of this evaluation should be viewed as indicative and approximate; they are the basis for a next quantitative step in which the performance of the indicators is tested against a data set from Dutch transitional and coastal waters.

[2] Optimal means: ecologically relevant, a relatively good scientific basis, good correlation with human pressures, easy to calculate, understandable by water managers and policy makers,.

2 Macrozoobenthos: structure, function, pressures and impacts

2.1 Definition of macrozoobenthos in the WFD

In this chapter, a short description is given of important characteristics of macrozoobenthos in transitional and marine waters, human impact factors (pressures) and how this (very roughly) affects structure and function of the macrozoobenthic community[3]. Here, explicitly is referred to macrozoobenthos and not the benthic ecosystem. The WFD only mentions 'benthic invertebrate fauna'. In the background document (Knoben & Kamsma 2004) this is not further specified. This background document refers to the much broader term 'macrofauna'. In follow-up documents the term 'macrofauna' is also used repeatedly (e.g. Van der Molen & Pot 2006). In further explanations on pressures in Van der Molen & Pot (2006) the term 'macrobenthos' is used, but no further description is given. Habitat-forming organisms such as mussels are also taken into account. Why meiozoobenthos and the microbial community are excluded is not explained, although these groups are quite important for the structure and functioning of the benthic ecosystem. Likely, smaller species (and often the younger specimens of the macrobenthos) are excluded because species identification is difficult, and not many studies on their relations with driving forces or feeding relationships exist. The larger (and less abundant) infauna also lack attention, mainly due to undersampling. The same holds for the larger and mobile invertebrate epifaunal organisms such as brittle stars and hermit crabs. Epibenthos can be sampled quantitatively with a benthic sledge, but this is not commonly done. Therefore, most benthic samples carried out with a box core or a multicore give an underestimation of the density and diversity of the epibenthos (and larger infauna). It must be noted here that epibenthic scavenger and predator invertebrates (e.g. crustaceans, gastropods and echinoderms) are strongly affected by bottom trawl fisheries (Rumohr 2000), since these groups thrive on the increased amounts of dead organisms and fish offal caused by fisheries. It might make sense to include this species group in further assessments of the benthic system.

Therefore it is mentioned explicitly that when referring to macrobenthos or benthos, infaunal macrozoobenthos and epifauna including biogenic habitat-forming organisms is meant (we will refer to macrobenthos in the rest of the document). This is commonly the fraction of infaunal and sessile epifaunal invertebrates that is sampled representatively from soft sediments such as sands and silty sediments with a box core or multicore, and retained at a sieve with a square sieve opening of 1x1 mm.

2.2 Pressures and impacts on benthic communities

Variability in environmental factors and ecological relationships cause variability in states of populations, communities and ecosystems. Many (human) pressures cause deviations from these 'natural states' of the ecosystem. The main environmental parameters (state variables) in the benthic environment are:

[3] The term "community" is chosen since literature commonly uses it when referring to macrobenthos. However, the term "cluster" or "assemblage" might be more appropriate when it concerns soft-bottom benthic life. Communities as usually sets of species that are tightly linked to specific environmental variables and ecological relationships. This link appears not very strong in benthic ecology, although there are clearly habitat preferences. Plant communities usually have relatively distinct borders, while (soft-bottom) benthic communities show a much more gradual change in species composition. Although this is a matter of proper definition, we decided to continue using the tan 'community'.

- Salinity
- Littoral or sublittoral height /depth and morphology
- Nutrients
- Water flow velocity and turbulence
- Soft substrate composition (mud content, organic matter content, median grain size)
- Soft and hard substrate elements
- Temperature
- pH

The main human pressures in the Dutch transitional and coastal waters are (not in any specific order):

- Eutrophication, leading to surplus deposition of organic matter and oxygen lack
- Pollution by metals and organics
- Coastal reconstructions affecting morphology, currents, substrate composition, adding hard substrate
- Sand extraction and dredging affecting morphology, currents, substrate composition
- Bottom-disturbing fisheries affecting morphology, currents, substrate composition
- Dumping and coastal nourishment affecting morphology, currents, substrate composition
- Climate change affecting temperature and pH

Salinity is a major natural state variable of a benthic habitat. In an estuary, strongly fluctuating salinities due to tidal variations strongly limit the number of species which can tolerate these variations. Therefore, in the monitoring and assessment of benthic habitats usually the following salinity zones are discriminated: oligohaline, mesohaline, polyhaline and euhaline.

The presence of nutrients, nitrogen and phosphorus, is of great importance for the primary production and abundance of phytoplankton in a water body, and consequently for the food availability for phytoplankton/detritus feeding benthos. It has been demonstrated that the total benthic biomass is approximately linearly related to the total primary production in a water body (Herman et al. 1999). This strong relationship, in combination with the strongly increased N and P concentrations in water bodies since the 1960s due to agriculture (eutrophication), has caused surplus deposition of organic matter which led to hypoxic or even anoxic events and changes or death of benthic communities in specific areas such as the Wadden Sea, the German Bight and fjords. No occurrences of hypoxia have been registered in the Dutch coastal waters.

The bottom height/depth determines if a bottom is permanently submerged (sublittoral) or periodically above the water level (littoral). Littoral conditions are more harsh than sublittoral conditions, and therefore limit the number of (often tolerant) species which can survive these conditions. Morphological features on small to large scales determine composition of substrate and benthic communities. Littoral or sublittoral conditions are therefore commonly used to discriminate benthic habitats. Extraction, dredging, dumping and nourishment all affect these features and thus benthic life significantly.

The water flow velocity is an important habitat factor, because (a) small benthic species can be flushed away at high flow velocities (as indicated by the shear stress), (b) it determines the sedimentation of inorganic and organic material, and (c) it influences feeding behaviour of the benthos. The practical use of water flow velocity as a parameter is not always easy. Therefore, the sand/mud content, which covariates with the water flow velocity, is usually used for this habitat factor. The use of sediment type (median grain size, silt content and organic matter content) is common in habitat classifications.

Temperature is mostly a natural variable which may have a strong impact on the species abundances. For example, strong winters may lead to a large reduction of species abundances in the following season. These large variations in abundances are one of the principal problems in the use of abundance indicators for the WFD assessment. Furthermore, climate change affecting water temperatures may lead to gradual shifts in abundances of temperature-sensitive species. This seems e.g. to be the case for the Baltic clam (*Macoma balthica*) in the Wadden sea. Lately, also pH changes in the water due to climate change appear to affect benthic life in temperate waters, such as echinoderms (Moulin et al. 2011).

In Dutch marine waters, bottoms are predominantly sandy bottoms. The presence of epibenthic bivalves bivalve beds may constitute small hard elements in a habitat, on which hard bottom species may occur. Due to this biogenic habitat, shear stress and turbulence is different from the surroundings, which influences settling (recruitment) of small inorganic and organic particles and organisms such as benthic larvae. Mussel beds in the Wadden Sea are therefore relatively rich in species that do not occur in the sandy sediment. These hard bottom species are also sampled and analysed in soft bottom aimed sampling procedures to some extent, partly because mussel and oyster beds constitute hard substrate structures, on which hard substrate species occur.

Bottom-disturbing fisheries have a severe effect on especially the epifauna, but also on the infauna in the top layers of the sediment. A large part of the epifauna and shallow benthos may be damaged or killed during a fishing track (Bergman & van Santbrink, 2000).

Sand extraction, dredging and nourishment may have locally large effects on the abundance and composition of benthic communities in sandy habitats. Surprisingly, Muxika et al. (2005) and Birklund & Wijsman (2005) have shown that sand extraction does not result in a larger fraction of opportunistic species. It appears that a benthic community is often not completely destroyed during sand extraction, and that apparently the community has sufficient resilience to maintain some balance between sensitive and opportunistic species.

In Table 2.1 an overview is given of the most important state parameters, their links to natural and human pressures, relevance in Dutch waters and possible indices or metrics.

Table 2.1: Overview of human pressures and natural factors in Dutch water bodies

State variable	Human pressure	Natural factor	Relevance in Dutch waters	Effects	Potentially suitable indicators	Reference
Organic material	Mostly sewage	Detritus (primary productivity, sedimentation regime)	Mostly no; has been in Wadden Sea; maybe to some extent in Eastern part of Westerschelde	Decrease of oxygen concentration and increase of micropollutants, leading to poorer growth and more opportunistic species.	The AMBI is useful to indicate this pressure, combined with the pressure micropollutants	Borja et al. (2000)
Micropollutants (heavy metals /POPs)	Uniquely human; associated with organic material and sludge.		Localised for heavy metals (copper, tin). Diffuse for organic pollutants	Imposex, bio-accumulation	AMBI.	Aakerman et al. (2004)
Nutrients (Si, N, P)	Agriculture (N, P)	Mineralization, mixing (no new nutrients added), riverine input (erosion)	Yes, coastal and transitional waters but decreasing ((P more strongly than N)	Increased algal blooms, nuisance plankton	The WFD phytoplankton metric is most suited for this pressure.	Loewe (2009)
Inorganic sedimentation	Dredging/dumping/nourishment, sand extraction	Wind, tidal currents, storms. The magnitude of this natural pressure is quite high along the Dutch coast	Extraction, dredging and dumping extensive along Dutch coast and in transitional waters	Smothering benthos, clogging of filter feeders, change community from filter feeders to subsurface	Infaunal trophic index, selective AMBI tuned to inorganic sedimentation.	Birklund & Wijsman (2005), Devlin et al. (2008)

State variable	Human pressure	Natural factor	Relevance in Dutch waters	Effects	Potentially suitable indicators	Reference
				deposit feeders		
Physical disturbance	Bottom trawl fisheries fishing, dredging/dumping, sand extraction	Wind, tidal currents	Extensive	Benthos mortality, reduction diversity, Change community structure	For bottom trawl fisheries a selective AMBI is currently developed. For dredging and sand extraction abundance indicators are useful; Existing AMBI is not useful because no increase of opportunistic species occurs.	Lindeboom & De Groot (1998), Muxika et al. (2005)
Climatic changes (both trend and cycles)	CO_2 increase, temperature rise, acidification	El Niño, NAO	Yes	Change community structure	Not yet clear.	Tsimplis et al. (2006), Moulin et al. (2011)
Salinity	Pulse-wise, unnatural variations in freshwater input in transitions (Waddenzee) and near-coastal waters (Voordelta) due to dams and locks	Some estuaries (Eems-Dollard, Westerschelde) still receive fairly natural input of fresh water and tidal currents	Yes	Benthos mortality, reduction diversity, Change community structure	Density or biomass. Indicator species sensitive for salinity shocks, e.g. *Cerastoderma edule*.	Steenbergen (2004)
Invasive species	Often by human influences, accidental (e.g. *Ensis directus*)	(e.g. introduction of Ensis by bilge water, introduction of Japanese oyster)	Yes	Change community structure	Abundance of specific exotic species. Invasive species index is possible.	Brinkman & Jansen (2007)

2.3 Variability of macrobenthic communities

Benthic species are grouped in communities that are structured by habitat variables like salinity, substrate (soft bottom/hard bottom), sediment composition (muddy/sandy), depth (subtidal or intertidal), current velocity/shear stress, temperature and primary production of phytoplankton. No in-depth description is given here of the composition of macrobenthos in transitional and coastal waters. Such descriptions can be found in Rees et al. (2007) and Borja et al. (2000). However, various features of the benthos are of high relevance for setting up indices.

First, spatial variation of the macrozoobenthos composition is large. In the North Sea, ecotopes, 'étages', and communities (Van Hoeij 2004, Borja 2009) can be discerned, which all display a more or less constant composition within a certain area (Rees et al. 2007). Such lower levels of variation are commonly associated with the physical and biological characteristics of this specific area. Examples are submerged sandbanks. They commonly show a different benthic community composition than surrounding 'flat' areas. Within such sandbank areas, often multiple different communities can be discerned (Craeymeersch et al. 1990). However, one should realise that such communities are not clearly discernable. Such communities in soft sediments are often better characterised by the term assemblages. Boundaries between communities usually are not very distinct, and consist of gradients in species composition. Nevertheless, multivariate statistical analyses commonly depict such clusters of organisms and assessing major clusters, communities or assemblages is a practical approach to assess spatial variability in species distribution and to couple such assemblages to environmental variability. In the Dutch BEQI method the spatial variability of communities is explicitly taken into account by pooling several samples within an ecotope-year.

Another major aspect is the temporal variability of the benthic community. The composition of the benthic community often is assumed to be more or less constant at a particular site. However, dynamics of sedimentation, resuspension and recruitment on coastal areas often cause an interannual variability of the benthic abundance, biomass and composition. This is why sampling for monitoring purposes often takes place in autumn, after a relatively long period with low climatic dynamics and at the end of the benthic recruitment processes leading to a relatively stable community composition. However, temporal variability can be substantial (Armonies 2000) and this is an issue that should be taken into account when setting up indicators and a monitoring plan.

2.4 Pressures, states and impact: DPSIR

In Heink & Kowarik (2010) an extensive overview is presented of the various definitions or uses given to the term "indicator" so far[4]. From their analysis, they suggest the following broad definition for the term indicator:

An indicator in ecology and environmental planning is a component or a measure of environmentally relevant phenomena used to depict or evaluate environmental conditions or changes or to set environmental goals. Environmentally relevant phenomena are pressures, states, and responses as defined by the OECD (2003).

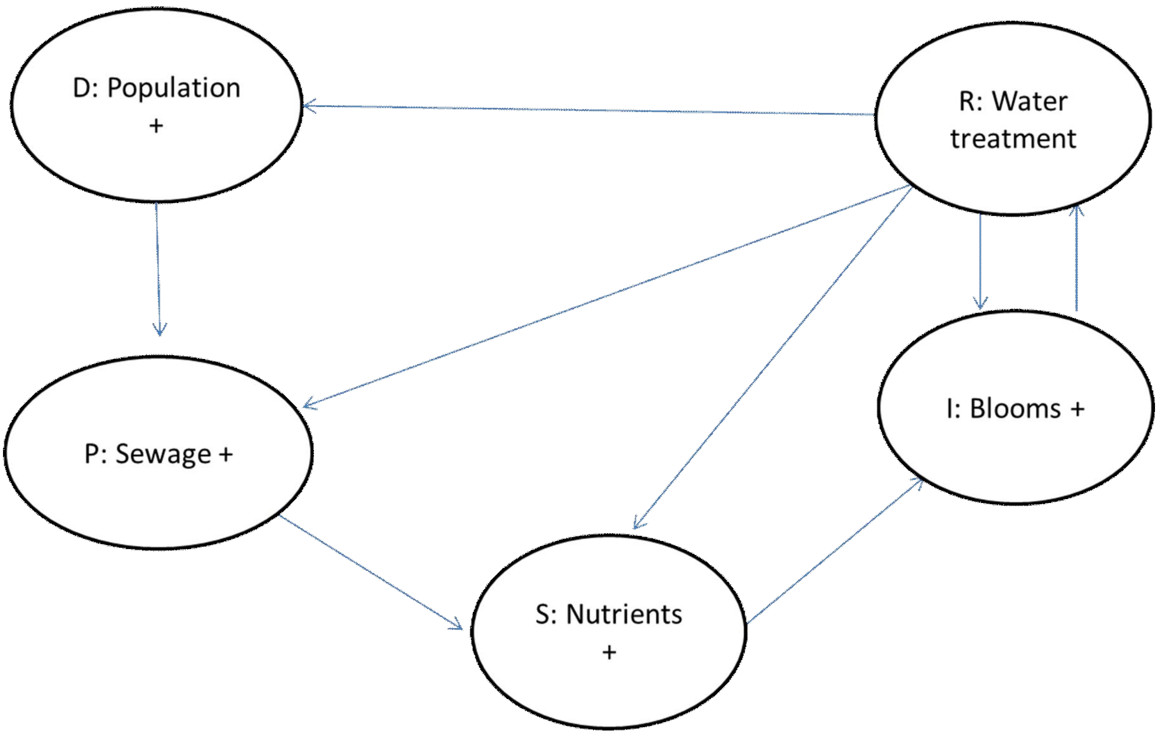
Hence, they clearly couple the use of the indicator to the PSR (pressure state response) model developed by the OECD, from which the European Environmental Agency (EEA)

[4] *An index can be defined as an integrated and/or complex set of indicators.*

further developed the DPSIR model (Smeets & Weterings 1999). According to the DPSIR framework there is a chain of causal links starting with ‘Driving forces’ (economic sectors, human activities) through ‘Pressures’ (emissions, waste) to ‘States’ (physical, chemical and biological) and ‘Impacts’ on ecosystems, human health and functions, eventually leading to political ‘Responses’ (prioritisation, target setting, indicators), which may be linked back to Drivers and/or Pressures.

An example is population increase (D), an increase in sewage input in coastal waters (P), a rise in nutrient concentrations (S), and as a consequence algal blooms and oxygen deficiency and dangers to human health in near-bottom waters and sediments. The response (R) would then be wastewater treatment. A schematic overview of these relations is given in the figure below. Setting up indicators also belongs to the response, in the view of Smeets & Weterings (1999). Although the DPSIR is mainly set up for finding indicators of human-induced pressures, it could also be applied to drivers and pressures from a natural source, such as the NAO (North Atlantic Oscillation) or El Niño, which cause changes in the Atlantic warm water current patterns or climatic (rain, wind) patterns, respectively.

Figure 2.1: Illustration of the DPSIR model for sewage



This means that for each Driver/Pressure combination the step to the corresponding “State”, “Impact” and “Response” needs to be taken. Indicators that need to express a certain State of a variable or Impact on an ecosystem necessarily need to comply to these same rules, i.e. the indicators of a specific DPSIR chain need to be related as well. Of course, this schematic is indicative, not absolute. A driver may have a direct effect on a ‘state’ or ‘impact’ as well.

In order to evaluate the indices or their components, it is therefore important to check whether they followed a procedure of relating pressures and impacts as described in DPSIR. Note that DPSIR is just a name coined to a logical method of coupling cause and effects of human influence on ecosystems or their parts, and a way to measure and manage this influence to mitigate or compensate for the unwanted effects. It is defendable that this method makes sense both logically and scientifically. Setting up a method from scratch would likely end up with quite a similar approach. In this report, the emphasis is clearly on the P-S-I part of the DPSIR chain.

Recently, Borja et al. (2011) has published a straightforward method to validate and quantify the response of indicators and metrics to a pressure index. The pressure index is calculated as the average of specific pressures and their estimated intensities (ranging from 1 to 3). This pressure index appears to correlate reasonably well with many indicators and metrics, which in general supports the application of these benthic assessment tools. More specifically, this method can be used to validate and compare the performance of related indicators and metrics, leading to an evidence-based selection of them instead of only theoretical considerations. It is recommended to check if other methods of adding up human pressures may give further improved results.

3 Background on WFD and selection of relevant benthic indicators

As a prerequisite for the Water Framework Directive (WFD) various multi-metrics containing several indicators have been and are being developed to give a measure of the ecological state of the benthic ecosystem in coastal and transitional waters as a reaction to human pressures. These indices are commonly based on quantitative calculations based on species composition, abundance (density, biomass) data and species sensitivity data (AMBI, BQI). Assigning species or species groups and their abundance to specific pressures and assigning them an EQR value is a difficult task and often lacks scrutiny or depth in many studies. This often hinders the more general applicability of such indices, and limits them to specific geographic areas and/or pressures.

3.1 The WFD and the benthic quality assessment

In the WFD context, the state of an ecosystem needs to be fit into 5 classes, ranging from ‘high (reference state) to ‘poor’ (maximum disturbance). In Annex V, table 1.2, a more precise description of the ‘best’ three classes of ecological status of the benthos, the so-called “normative definition”, is given for transitional waters. This clearly depicts the features of the benthic ecosystem that need to be assessed.

Table 3.1: WFD normative definitions for benthos in marine waters

<table>
<tr><th>Status</th><th>Normative definition</th><th>Comments</th></tr>
<tr><td>High status</td><td>The level of diversity and abundance of invertebrate taxa is within the range normally associated with undisturbed conditions.

All the disturbance-sensitive taxa associated with
undisturbed conditions are present.</td><td rowspan="2">Diversity is commonly translated into the indicator “species richness”.
Abundance in the strict sense gives a calibration problem of the indicator “abundance” since there is no clear relationship between abundance and ecological status (see Pearson Rosenberg 1984 and the effect of eutrophication). Therefore, many NEA GIG countries seem to have translated “abundance” into a “diversity” indicator which uses relative abundances, such as a Shannon index or a Simpson index.

Pollution is only a single type of human pressure. Therefore, it is common practice to interpret “pollution” as “human pressure”.</td></tr>
<tr><td>Good status</td><td>The level of diversity and abundance of invertebrate taxa is slightly outside the range associated with the type-specific conditions.

Most of the sensitive taxa of the type-specific communities are present.</td></tr>
</table>

Status	Normative definition	Comments
Moderate status	The level of diversity and abundance is moderately outside the range associated with the type-specific conditions. Taxa indicative of pollution are present Many of the sensitive taxa of the type-specific communities are absent	An indicator for the classification of taxa being sensitive and tolerant in relation to human pressures, is required. It is probably important to focus on the major human pressure(s) on the benthic community present in the specific water body. Therefore, in principle three indicators are at least required: 1. Diversity 2. Abundance 3. Sensitive and opportunistic species

In these normative definitions, the reference state is not defined in any structural or functional ecological terms. According to the WFD, it is “simply” an ecosystem that functions with a minimum of human influence[5]. The fact that it is defined in this way means that a very explicit choice has been made for defining the reference situation: a pristine ecosystem status in some period and/or place where man’s influence is negligible or absent. However, such pristine conditions are not well defined, do not consist of one typical state and data about such conditions are rare.

For the benthic ecosystem this means that indicators need to be sought that reflect the presence of human pressures, clearly discriminate them from natural pressures/variation, and “hindcast” a reference state or good state from studies that relate these specific indicators quantitatively to the various human and natural pressures (thus finding a ‘reference level’). In practice, it is often difficult to find datasets that describe the reference situation, and benchmark data reflecting the highest known quality of benthic communities are often used.

3.2 Selection of relevant indicators

In the tables below, an overview is given of the benthic metrics currently used in EU countries in the WFD context (Table 3.2), and other indicators used in (marine) ecology for assessing benthic ecosystem status or functioning (Table 3.3). On the basis of the developed metrics in the EU, a choice is made of the relevant indicators to be discussed and reviewed. Partly, the relevant indicators come from the metrics that have been developed for use in the WFD.

Table 3.2 has been translated and adapted from Ysebaert et al. (2008).

[5] *There are no, or only very minor, anthropogenic alterations to the values of the physico-chemical and hydromorphological quality elements for the surface water body type from those normally associated with that type under undisturbed conditions. The values of the biological quality elements for the surface water body reflect those normally associated with that type under undisturbed conditions, and show no, or only very minor, evidence of distortion. These are the type-specific conditions and communities.*

Table 3.2: Overview of the WFD metrics for marine benthos in Western Europe

Member country	**Metric**	**Indicators used in the metric**				**References**
		Sensitive/ opportunistic species	*Diversity*		*Abundance*	
Spain France	m-AMBI	AMBI	Species number	Shannon-Wiener Index		e.g. Borja et al. (2011)
Germany	Marbit	Sensitive/ opportunistic species	Taxonomic Spread Index		Density distribution	Meyer et al. (2007)
Germany	AeTV	AeTI	Average species number	ADF		Krieg (2010)
United Kingdom	IQI	AMBI	Taxa number	Simpson index	Density	WFD-UKTAG (2008)
Denmark	DKI	AMBI	Species number	Shannon-Wiener Index	Density	Josefson et al. (2009)
Norway	NQI	AMBI	SN Index	Shannon-Wiener Index	Density	Josefson et al. (2009)
Portugal	P-BAT	AMBI	Species number			Pinto et al. (2009)
Sweden	BQI	$ES50_{0.05}$	Species number		Density	Rosenberg et al. (2004)
Nether-lands	BEQI	Bray-Curtis	Species number		Density Biomass	Van Hoey et al. (2007)
Belgium	BEQI	Bray-Curtis	Species number		Density Biomass	Van Hoey et al. (2007)

See the Abbreviations list for explanations.

If we make a frequency analysis of the methods used by the 9 NEA-GIG countries, the following results appear in Table 3.3.

Table 3.3: Use of indicators in the Western European marine benthos metrics

Method	Nr of methods	Remarks
Species number	9	The species number is also used to compensate for undersampled areas.
Shannon index	3	
Simpson index	1	
ADF diversity index	1	5 metrics use a diversity index
Density	6	Density is often used to correct for undersampled areas
Biomass	1	Only the BEQI contains biomass
AMBI	5	The most popular sensitive/opportunistic species index
Bray-Curtis	1	Only in BEQI
ES50	1	Only in BQI
AeTI	1	Only in AETV

Another set of indicators are those commonly used in ecology for e.g. species diversity, or functional diversity, but have not been used in the recently developed WFD metrics. These are mentioned in Table 3.4 below. These indicators are also evaluated in this study.

Table 3.4: Additional indicators which are reviewed in this report.

Indices/Metrics	Description	References
Diversity		
Evenness E	The relative abundance of the different species making up the richness of an area	Schroeder (2003)
Species-abundance plots	Species are ranked in sequence from most to least abundant along the horizontal (or x) axis	Warwick & Clarke (1996)
Abundance-Biomass Comparison	Species are ranked in order of importance in terms of abundance or biomass on the x-axis (logarithmic scale) with percentage dominance on the y-axis (cumulative scale)	Meire & Dereu (1990)
Sanders/Hurlbert no.	The estimated number of species among 100 individuals (ES100)	Reiss & Kroencke (2005)
Collector's curve	Expresses the number of species as a function of the number of specimens caught	Magurran (1988)
Sensitive/opportunistic species		
BOPA	The opportunistic polychaete/amphipod ratio	Dauvin & Ruellet (2007)
ITI	Based on feeding types	Krieg (2010)
r/K strategist	Based on survival strategy	Lavaleye (1999)
Abundance		
Density	Number of individuals per m^2	Lavaleye (1999)
Biomass	Weight of benthos per m^2	Duineveld et al. (2007)
Size spectra	Size-abundance distribution	Duineveld et al. (2007)

4 Evaluation of benthic indicators and metrics

4.1 Introduction and evaluation criteria

Species and their populations can be characterized in several ways, such as:
- taxonomic identification: species, genus, family, order, etc
- age and size
- density and biomass
- feeding route(s)
- food sources
- sensitivity/tolerance for human and natural pressures

Species can roughly be classified as commonly occurring and rare species (Gittenberger, 2011). By definition, the common species dominate the structure and functioning of the macrobenthic community and must at least be assessed adequately. For example, the Shannon-Wiener diversity index accounts for the density of species, giving less weight to less abundant species.

The indicators and metrics usually have been used or are developed in a specific context and with a specific goal. AMBI and BQI are based on the Pearson-Rosenberg model for benthic disturbance by oxygen depletion from 1978 (Borja et al. 2000, Rosenberg et al. 2004).

Indicators and metrics often lack scrutiny regarding calibration to (human) pressures and community state variables, the typical steps asked for in the DPSIR approach (or any other applicable approach). Two important aspects of any indicator/metric are the pressure-impact correlation and the spatial and temporal variability.
Firstly, the causal and quantitative or correlative relation of the indicator/metric with the pressure and with the state or system variable. A recent study of Borja et al. (2011) is a nice example of how these correlations can be investigated using a so-called pressure index. Secondly, the spatial and temporal aspects of the indices are major components in the calibration of indices and indicators in general. The quality of calibration depends much on taking these variation sources into account. Ideally, one should couple a high-resolution view of human pressures and environmental variables with a comparable view of the indicators/ metric to get a proper understanding of the performance of the indicator with what it is supposed to indicate. Here it needs to be stressed that such relations should be studied to a background of natural drivers and pressures. Ideally, an indicator is sensitive and selective for particular pressures, e.g. organic material/pollutants, fisheries or sedimentation, for which it has been designed.

In theory, indicators can be set up on a single-species level, a species group level or on the entire community level. An example of a single species indicator is the OSPAR ecological quality objective “dog whelk” (*Nucella lapillus*) for tributyltin (Gmelig Meyling 2006). It can be considered to be a very useful data exploration to routinely analyse time-series of the dominant benthic species, and specific indicator species , as a part of the WFD assessment, and try to explain possible trends. This will give a rough understanding of the developments in the benthic community and thus a context to interpret index values.
The BOPA (Dauvin & Ruellet 2007, see section on disturbance indicators below) is a nice example of a species group indicator which considers the total number of individuals collected in the samples, the frequency of opportunistic polychaetes, and the frequency of amphipods.

The AMBI index (Borja et al. 2000) using groups of sensitive to opportunistic species[6] is another example of a species group approach. Community level indicators describe the general status (biomass, number) of all species in a sample. Multivariate statistics commonly do this through a cluster or ordinate analysis, such as a Bray-Curtis dissimilarity index, MDS, TwinSpan, DCA, etc. (Jongman 1995, Warwick & Clark 1996, Van Hoey et al. 2007).
WFD metrics for marine benthos in most cases contain a suite of e.g. three indices, presumably because the WFD normative definitions require three indicators for diversity, abundance and sensitive/tolerant species. In order to evaluate the available WFD metrics, they are broken down into their underlying indices and indicators, which are evaluated separately. In paragraph 4.1 the diversity indicators are evaluated, in paragraph 4.2 the abundance indicators, in paragraph 4.3 the sensitive/opportunistic species indicators and in paragraph 4.4 the composite metrics are evaluated. In paragraph 4.5 the conclusions and recommendations for all the evaluation paragraphs are given.

The evaluation of the indicators will be performed using the following evaluation criteria:

a) Indicator design. How does the indicator work? Which question does it try to answer? Is the indicator ecologically sound?
b) WFD compliance. Does the indicator comply with the WFD normative definitions for marine benthos? This is a knock out criterion; if an indicator is not WFD compliant it cannot be used for a WFD metric.
c) Practical aspects. Is the indicator user friendly? What are the data and calculation requirements?
d) Sample area sensitivity. How sensitive is an indicator for the sampling surface area?
e) Indicator calibration. Does the indicator need species reference data? For example to calculate the AMBI it needs species reference data for undisturbed (baseline) status and maximum disturbed status, but the Shannon index only needs basic species and density data.
f) Human pressure calibration (sensitivity and correlation). Has it been demonstrated with human pressure/indicator response curves that the indicator is sensitive to human pressures? How good is the correlation of this relationship? See Borja et al. (2011) for a useful method for human pressure calibration.
g) EQR normalization . How is the indicator response fitted and scaled to an EQR? How reliable is this normalization method? Note: two point calibration (using a reference point EQR = 1 and a bad point with a EQR = 0) and Swedish three point calibration (with the observed Good-Moderate boundary as additional point) are the two most common normalization methods. The Swedish method is probably ecologically the best normalization method, but requires a good calibration of the pressure-response scale which if often not available. Two point normalization is clearly a simplification of the ecological reality, but is a pragmatic method which also appears to work in the Intercalibration exercise. Two point normalization usually employs a linear curve and equidistant EQR scale.
h) Intercalibration. Can the indicator be intercalibrated? Although this criterion is not meant as a quality criterion for the indicator itself, it is a measure for comparability with other countries and as such an important characteristic.
i) Conclusions and average score of the indicator

---

[6] Sensitive or tolerant to a specific type of change in a state variable or introduced pollution. Although Borja and co-workers (and in many other studies) have been using such lists of species extensively (Borja et al. 2000), only lately it has been studied to which specific parameters the supposed sensitivity or tolerance relates (Borja et al. 2011).

The evaluation results per criterion are first described, and at the end summarized in a conclusion. It has to be stressed here, that the evaluation is partly based on literature, and partly on the authors' expert judgement. It is used for intercomparison of the indicators and not for absolute performance. The scores mentioned in the evaluation tables are the averages of the scores by the authors. They have scored the indicators independently. Discussion is possible about the weighting of the criteria. The authors acknowledge that calibration with human pressures is a crucial evaluation criterion. An indicator performing below 3 on pressure calibration is not considered a serious candidate at this moment. If this is due to lack of knowledge, additional studies are needed. If it is due to lack of performance, the parameter is not likely to be a good indicator ever. If an indicator scores low on pressure calibration due to lack of performance, this will be marked with red colouring.

Please note: in order to quantify the evaluation, the following guidelines are used for the scoring of the criteria:

5: proven good OR not necessary/relevant for this indicator/metric
4: proven reasonably good OR probably good
3: proven moderate OR probably reasonably good
2: proven inadequate OR probably moderate
1: proven bad OR probably inadequate?

4.2 Evaluation of diversity indicators

Species diversity is the simplest of all biodiversity measures, and applicable as a measure for single samples. The most common metrics for species diversity are the Species Richness, the Simpson index, the Shannon-Wiener index and Evenness (Vos 2006).

There are various "levels" of diversity. Commonly, diversity is the sample diversity, also known as alpha diversity. Beta diversity is the amount of variation in a collection of sample units, and gamma diversity is the overall diversity in a collection of samples, usually at a "landscape" level. The diversity discussed here concerns sample and ecotope diversity, the alpha and beta level diversity.

Measuring the number of species and looking at their density, any benthic environment will show a large intra-annual variation due to recruitment processes and seasonality in productivity. Also, climatic events such as a severe winter may strongly impact species abundances **(**Reiss & Kröncke 2005). Furthermore, it may be more difficult to identify species in a juvenile life stage in spring. It is generally accepted by Dutch marine benthic experts that sampling in autumn is the more favourable moment, since species are more full grown and can be identified more easily, and the abundances and biomasses of species are more stable and larger than in spring. Therefore, in the Dutch guidelines for WFD sampling and assessment (Faber et al. 2011) it is recommended to sample at least in autumn, and if necessary additionally in spring.

It has to be noted that the Intermediate Disturbance Hypothesis (IDH, Connell 1978) predicts that diversity will be maximized in communities experiencing relatively low levels of disturbance, see Fig. 4.1 below. Although the IHD has been refined considerably, the general view holds (Wilson 1994, Roxburgh et al. 2004). Borja et al. (2000) has shown similarly that at low AMBI-values (around 1) with a low level of disturbance the species richness and diversity are higher than in undisturbed situations (AMBI-value ca. 0), and that at higher disturbances than AMBI value 1 these two diversity parameters decrease. In conclusion, the relationship between pressure and diversity appears to be not entirely linear but when the pressure level is in the second part of the pressure-diversity curve then a useful correlation is expected to be present.

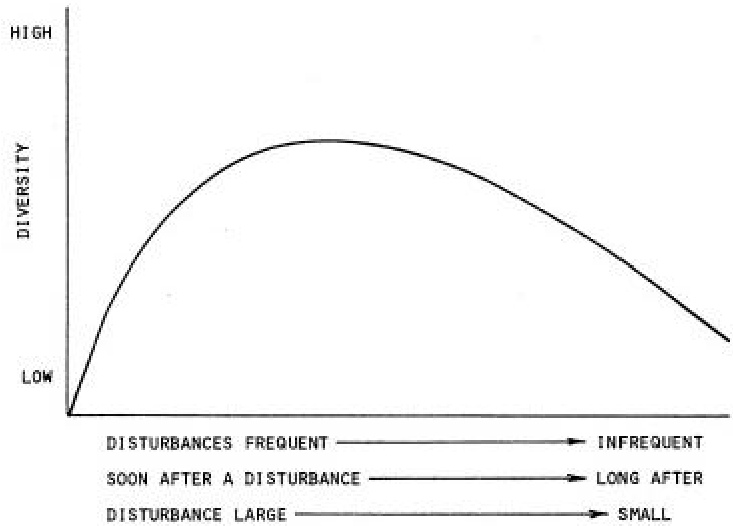
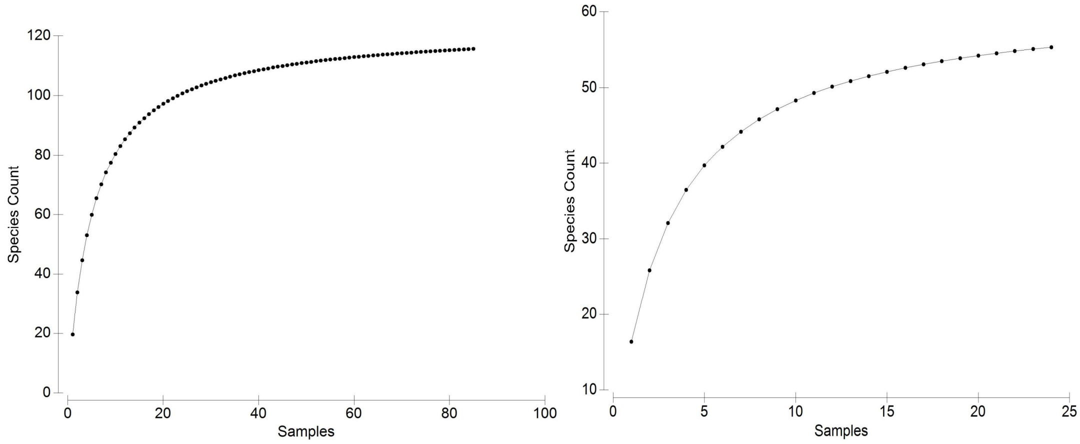


Fig 4.1: Schematic representation of the intermediary disturbance hypothesis (from Connell 1978).

4.2.1 Species Richness indicators

Species richness is the most basic diversity indicator, and an important and broadly used indicator for benthic quality. The number of species per sample or ecotope is a measure of richness. A disadvantage of this indicator is that it depends on the sampled surface area. Therefore, several methods have been developed to estimate the number of species in a standardized manner.
The basic method to examine the effect of the sampled area on the number of species found is the construction of a species accumulation curve (Van Hoeij et al. 2007). In this curve, the number of species found is plotted on the Y-axis in relation to the sampling area on the X-axis. In Figure 4.2 below, examples of species accumulation curves are given for two ecotopes in the Westerschelde. It is clear from these figures that (a) many samples have to be taken to approximate the asymptotic number of species, (b) in the polyhaline zone the number of species is much larger, in principle due to lower variations in salinity leading to a more favourable benthic environment.

Figure 4.2: species accumulation curves based on RWS-MWTL data for the Westerschelde ecotopes polyhaline intertidal sand (left) and mesohaline intertidal sand (right).



In order to obtain spatially sufficiently stable and reliable data, it was decided in the NEA GIG benthos working group to standardize the marine benthic assessment to a sampling area of 0.19 m^2 per ecotope (Borja et al. 2011).

There are several options to obtain the species richness from this curve:

a) use the collectors curve stop criterion
b) interpolate the curve to a standardized number of individuals
c) standardise to a sufficient and fixed sampling area

These three methods are evaluated below.

Species Richness (S) - Collectors curve

Evaluation criterion	Evaluation result	Score
Indicator design	Species Richness S – Collectors curve. The sampling can be stopped when doubling of the sampling area leads to less than a defined increase in the number of species, e.g. 1, showing that the asymptotic/true number of species has been reached. As more specimens are sampled, a collector's curve can reach an asymptotic value but they often don't due to the vague boundaries of ecotopes. Important ecological information; larger human pressures in general lead to less species. However, rare species contribute as much to this indicator as common species which is a disadvantage. Furthermore, the evenness of the density distribution is not accounted for.	3
WFD compliance	Yes, diversity indicator	5
Practical aspects	Very simple model. Large sampling areas have to be used in order to fulfil this criterion. This method therefore only is used in inventory studies of all species in an area, and not for routine monitoring.	2
Sample area sensitivity	Due to the large sampling area, S will stabilize to an nearly asymptotic value.	2

Evaluation criterion	Evaluation result	Score
Indicator calibration	The indicator does not have to be calibrated	5
Human pressure calibration	S is in general sensitive to pressures, but no selective human pressure information can be obtained.	3
EQR normalization	This indicator can be calibrated well, but close attention has to be paid to a standardized sample area.	5
Intercalibration	S is often used in NEA GIG metrics.	4
Conclusions and average score	Due to the large sample areas necessary, this method is too expensive for routine monitoring but only used in broad species surveys.	3.6

Species Richness - Estimated number of species (ES50)

Evaluation criterion	Evaluation result	Score
Indicator design	Species Richness – Expected number of species at 50 individuals (ES50). Sanders' rarefaction index (Sanders 1968), the number of species is calculated in relation to a certain number of individuals, e.g. from a sample 100 individuals may be collected at random and among these individuals the number of species can be identified and counted. This will give the estimated number of species among 100 individuals (ES100). Sanders' method of calculation over-estimated the number of species and, therefore Hurlbert (1971) later corrected the formula. Important ecological information; larger human pressures in general lead to less species. However, rare species contribute as much to this indicator as common species which is a disadvantage. Furthermore, the evenness of the density distribution is not accounted for.	3
WFD compliance	Yes, diversity indicator.	5
Practical aspects	A calculation step using several samples is necessary, which is a practical disadvantage.	3
Sample area sensitivity	This statistical standardization method is a good way to solve the sample area problem.	4
Indicator calibration	Has been calibrated (in BQI, Rosenberg). This indicator has to be calculated from data from a series of samples.	4
Human pressure calibration	S is in general sensitive to pressures, but not specific human pressure information can be obtained.	3
EQR normalization	The ES50 can be calibrated well, e.g. using the ratio of the assessment value divided by a reference value.	4
Intercalibration	ES50 is not often used in NEA GIG metrics.	4
Conclusions and average score	The ES50 is a relevant method to estimate the Species Richness.	3.8

Species Richness - Standardized sampling area

Evaluation criterion	Evaluation result	Score
Indicator design	Species Richness – Standardized sampling area. The number of species are counted using a standardized sampling area. This leads to values of S that can be compared in a meaningful way. However, rare species contribute as much to this indicator as common species which is a disadvantage. Furthermore, the evenness of the density distribution is not accounted for.	3
WFD compliance	Yes, diversity indicator.	5
Practical aspects	Samples have to be pooled to a standardized sample area per ecotope. E.g., in the NEA GIG intercalibration it has been agreed to standardize the total sample area per ecotope to 0.19 m^2.	3
Sample area sensitivity	Due to the sample area standardization, the sample area sensitivity has become lower, but remains a point of attention.	3
Indicator calibration	The indicator does not require calibration data.	5
Human pressure calibration	S is in general sensitive to human pressures, but no selective human pressure information can be obtained. In Borja et al. (2011) a reasonable human pressure sensitivity was observed, but the correlation was not very good (R^2 = 0.38).	3
EQR normalization	This indicator can be calibrated using two-point calibration and three point calibration.	5
Intercalibration	S is often used in NEA GIG metrics.	5
Conclusions and average score	The most straightforward operational method to assess the species richness of ecotopes.	4.0

4.2.2 Margaleff's d

Evaluation criterion	Evaluation result	Score
Indicator design	Margaleff's *d* is a measure of the number of species S corrected for the amount of individuals N found in the sample (pool): d = (S – 1)/log(N). However, rare species contribute as much to this indicator as common species which is a disadvantage. Furthermore, the evenness of the density distribution is not accounted for.	3
WFD compliance	Yes, a diversity indicator	5
Practical aspects	The indicator is simple to calculate. Density data are usually reliable.	4
Sample area sensitivity	Not sampling area sensitive: one sample suffice. This is an important advantage of *d* over the use of S.	4
Indicator calibration	The indicator does not has to be calibrated.	5
Human pressure calibration	S is in general sensitive to pressures, but not selective human pressure information can be obtained. Borja et al. (2011) found a relatively good correlation of *d* with a human pressure gradient (R^2 = 0.48), which was significantly higher than the correlation for S with the same pressure gradient (R^2 = 0.38).	3

Evaluation criterion	Evaluation result	Score
EQR normalization	This indicator can in principle be calibrated, e.g. using the ratio of the assessment value divided by a reference value.	5
Intercalibration	*d* is only used in the BAT.	4
Conclusions and average score	Due to the lower sensitivity to the sample area and higher human pressure correlation than S, *d* seems to be preferable over S.	4.1

4.2.3 Shannon Wiener index (H')

Evaluation criterion	Evaluation result	Score
Indicator design	The Shannon index is the most widely used diversity index. It is calculated as follows: $H' = \Sigma p_i * \log p_i$ in which p_i is the relative density (fraction) of species i. The log base can be 2, e or 10. In the NEA GIG metrics, log base 2 is used because it gives a good response to changes in relative abundances, while e.g. log base 10 gives weak responses (Borja et al. 2011). Therefore, log base 2 is generally recommended. The Shannon index gives a very informative picture of the number of species (S) combined with the evenness (E) of a (pool of) sample(s). Rare species have a low contribution to this index which is an advantage. Due to the incorporation of E, H' has a higher information value than S alone. If it is desired to discriminate between S and E, one of these two indicators must be added to H'.	5
WFD compliance	Yes; because it is a diversity indicator.	5
Practical aspects	H' is very simple to calculate.	5
Sample area sensitivity	This index is less sensitive for the sampling area than S	4
Indicator calibration	This indicator does not require calibration data.	5
Human pressure calibration	The Shannon index appeared to be sensitive to human pressures, and to correlate quite well (R^2 = 0.64) with a human pressure index gradient, and was the highest scoring indicator in a study of Borja et al. (2011). There is no good relation with fisheries or extraction/sedimentation.	4
EQR normalization	This indicator can be calibrated well, e.g. using the ratio of the assessment value divided by a reference value.	5
Intercalibration	H' is regularly used in the WFD metrics	5
Conclusions and average score	A commonly used indicator with a good sensitivity for and correlation with human pressures.	4.8

4.2.4 Simpson index (λ or lambda)

Evaluation criterion	Evaluation result	Score
Indicator design	A measure that accounts for both richness and relative density of each species is the Simpson's diversity index. It is calculated as follows:	4

Evaluation criterion	Evaluation result	Score
	$\lambda = \Sigma p_i^2$ in which p_i is the relative density (fraction) of species i. The Simpson index reacts somewhat more strongly to the relative density, and in this way gives less weight to rare species than the Shannon index. An interesting aspect of λ is that $1/\lambda$ = S x E, in which E is the evenness of the density distribution. Therefore H' has a higher information value than S alone. Furthermore, lambda is less dependent on the sampling area than S. If it is desired to discriminate between S and E, one of these two indicators must be added to lambda.	
WFD compliance	Yes; because it is a diversity indicator.	5
Practical aspects	λ is very simple to calculate.	5
Sample area sensitivity	This index is less sensitive for the sampling area than S	4
Indicator calibration	Lambda does not have to be calibrated.	5
Human pressure calibration	The correlation with a pressure index gradient has additionally been calculated for this review (with courtesy to Borja, 2011) and appears to be considerably lower (R^2 = 0.27) than for the Shannon index.	2
EQR normalization	Using suitable reference data, this index can be calibrated well.	5
Intercalibration	λ is only used in the IQI, and could correlate with other metrics less well due to the lower human pressure correlation.	3
Conclusions and average score	A commonly used indicator but with a considerably lower correlation with human pressures than the Shannon index.	4.3

4.2.5 Evenness (E)

Evaluation criterion	Evaluation result	Score
Indicator design	The Evenness E gives a measure of the evenness of the density distribution of the species of a community. The Evenness can simply be calculated from all common diversity indices, e.g. from the Shannon index leading to the Pielou index as J' = H'/log(S), or from the Simpson index as E = 1 / (λ * S). In principle more even density distributions represent a higher diversity and higher community health than distributions in which only a few species dominate and the other species are rare. Therefore the evenness of a community is in principle important ecological information. The information content of this indicator is not as high as of H' and λ, because these indicators contain information on both S and E. It appears from a recent draft report on benthic evenness in the Voordelta coastal region (Craeymeersch et al. 2010) that the values for E are relatively stable in several subareas, and therefore less discriminative, compared to the number of species.	3
WFD compliance	Yes, a diversity indicator.	5
Practical aspects	E is very simple to calculate.	5
Sample area sensitivity	It is estimated that this indicator is not very sensitive to the sample area.	3
Indicator calibration	The indicator does not have to be calibrated.	5

Evaluation criterion	Evaluation result	Score
Human pressure sensitivity	General, but no specific, pressure information. It appears from Voordelta data (Craeymeersch et al. 2010) that evenness data are not very discriminative.	2
EQR normalization	This indicator has an intrinsically useful, EQR-like, scale from 0 to 1.	5
Intercalibration	Since this indicator is not used in WFD metrics, and reacts differently than S, the intercalibration potential is estimated to be low.	2
Conclusions and average score	A relevant ecological indicator, but not as informative as H'.	3.8

4.2.6 K-dominance plot

Evaluation Criterion	Evaluation result	Score
Indicator design	In a k-dominance plot species are ranked in sequence from most to least abundant along the horizontal (or x) axis. Their cumulative abundances are typically displayed on the y-axis. The k-dominance plot is very informative about the biodiversity of the analysed sample (pool). When k-dominance curve is used for comparing the biodiversity between many habitats, it is called multiple k-dominance curves. K-dominance plots can be based on density or on biomass data. The plot gives a picture of both the number of species and the evenness of the sample (pool). It is expected that this information could correlate with other diversity indicators.	4
WFD compliance	Possibly yes but is not expressed in a number, which makes it hard to compare directly.	5
Practical aspects	Calculation using k-dominance plots are not standard and require an additional calculation effort.	3
Sample area sensitivity	This plot is more or less sensitive to the sample area	3
Indicator calibration	A k-dominance plot gives a significant amount of visual information. However, for calculation purposes this information must be transformed into a numerical value. A logical point to determine a numerical value for the k-dominance from a k-dominance plot seems to be the number of species at the 50% (median) cumulative density or biomass.	2
Human pressure calibration	This evenness presumably reacts to pressures, which can be visible in the k-dominance value. However, no discrimination between human and natural pressures can be made.	2
EQR normalization	Using suitable reference data, this index can probably be calibrated.	4
Intercalibration	This index is not used in the NEA GIG metrics, but may be intercalibrated as a diversity indicator	3
Conclusions and average score	Detailed visual diversity information, but the transformation of k-dominance curves in a numerical value is a practical disadvantage.	3.3

In Figures 4.3A and 4.3B below, two k-dominance plots are shown based on density and biomass data.

Figure 4.3A: k-dominance plots for several Westerschelde ecotopes based on density data.

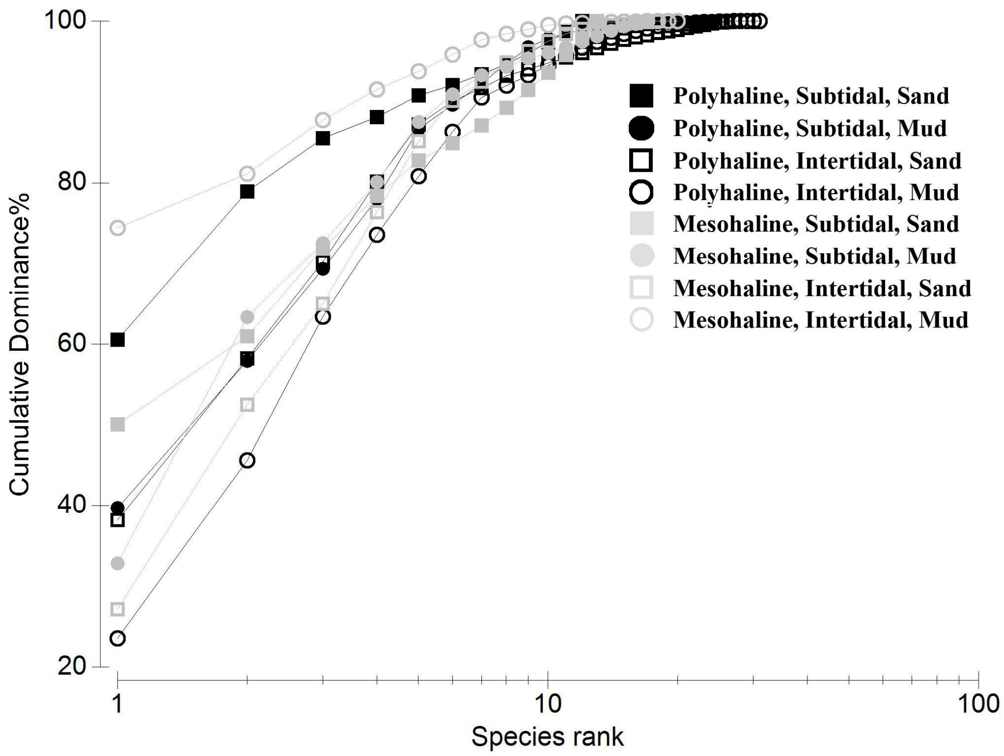
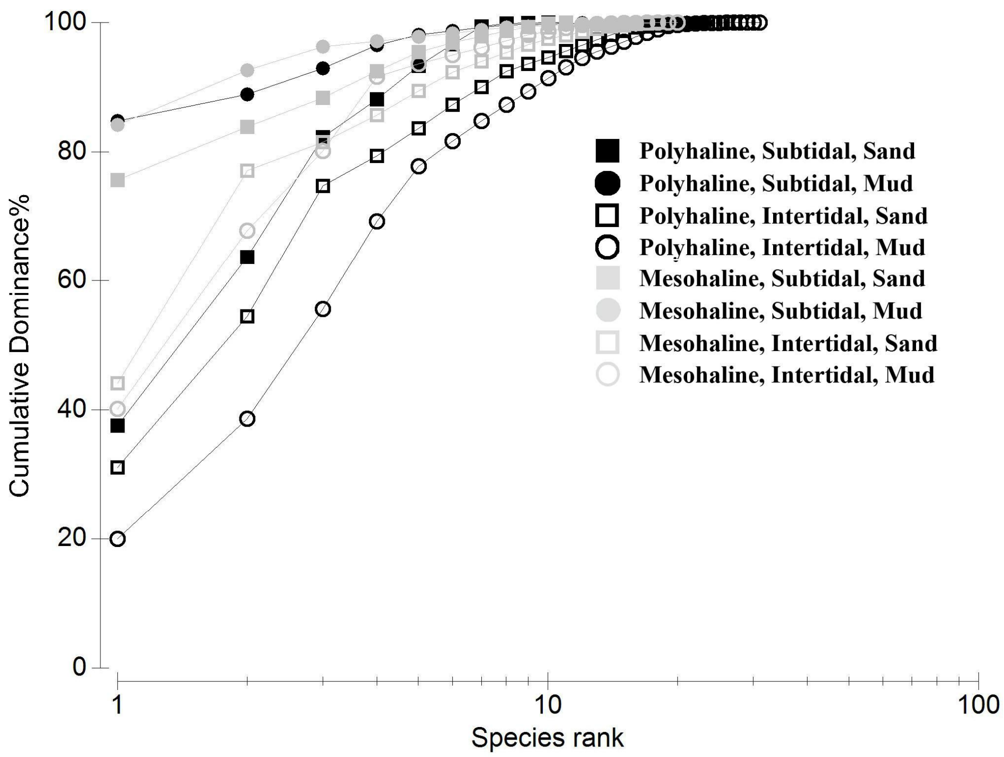


Figure 4.3B: k-dominance plots for several Westerschelde ecotopes based on biomass data.



4.3 Abundance, size and age indicators

4.3.1 Total density

Evaluation criterion	Evaluation result	Score
Indicator model	There are various ways in which NEA GIG countries have interpreted the demand for an abundance indicator in the WFD normative definitions. In the BEQI, total density per m^2 is calculated and assessed (Van Hoeij 2007, Twisk 2009). In various NEA GIG metrics, total density is used as a correction factor for insufficient sampling area. Also, in several metrics (such as the m-AMBI), relative densities are assessed in the form of a Shannon or Simpson index. Total density clearly is an important ecological indicator.	5
WFD compliance	Yes, abundance indicator.	5
Practical aspects	This indicator can be calculated very easily	5
Sample area sensitivity	Clearly, total density is very dependent on the sample area which is a disadvantage.	3
Indicator calibration	Has been calibrated. The calibration of total density in a WFD context is a problem. First, total density is especially related to the eutrophication status (Dekker, 1984). Second, according to the Pearson Rosenberg model, lower total densities may equally well indicate lower human pressures as a heavily polluted habitat (two sided distribution). Finally, large seasonal variations of densities of species may occur (Essink et al. 1986). These phenomena make it difficult to calibrate total density for the WFD compared with a natural reference on a historical time scale, which is necessary for the WFD. On a spatial time scale and with local pressures such as dredging, calibration seems possible (Ware 2009). Similar statements about the problems with calibration of total density have been made by Borja et al. (2000) and Meyer (2007).	3
Human pressure calibration	Total density is sensitive to many natural and anthropogenic pressures, such as seasonal temperatures, eutrophication, dredging and sand extraction. Density has clearly been related to the pressure dredging and dumping (Ware 2009). This pressure is clearly located in space. In time, density is especially an indicator for eutrophication and related phytoplankton status, and of seasonal temperature variations. It may not be easy to assign this indicator to specific human pressures. The use of the indicator density for the pressure eutrophication is not necessary and desired, since this pressure is assessed more precisely in the phytoplankton metric. Furthermore, natural seasonal variations of density may obscure structural changes of the community due to human pressures.	3
EQR normalization	This is possible, once the indicator has been calibrated adequately.	3
Intercalibration	Total density is not used in other NEA GIG metrics. This may lead to a lower intercalibration performance .	3
Conclusions and	The calibration problem of the indicator total density prevents its	3.8

Evaluation criterion	Evaluation result	Score
average score	application as an independent WFD indicator. For the assessment of localized pressures in space such as dredging, total density seems to be useful.	

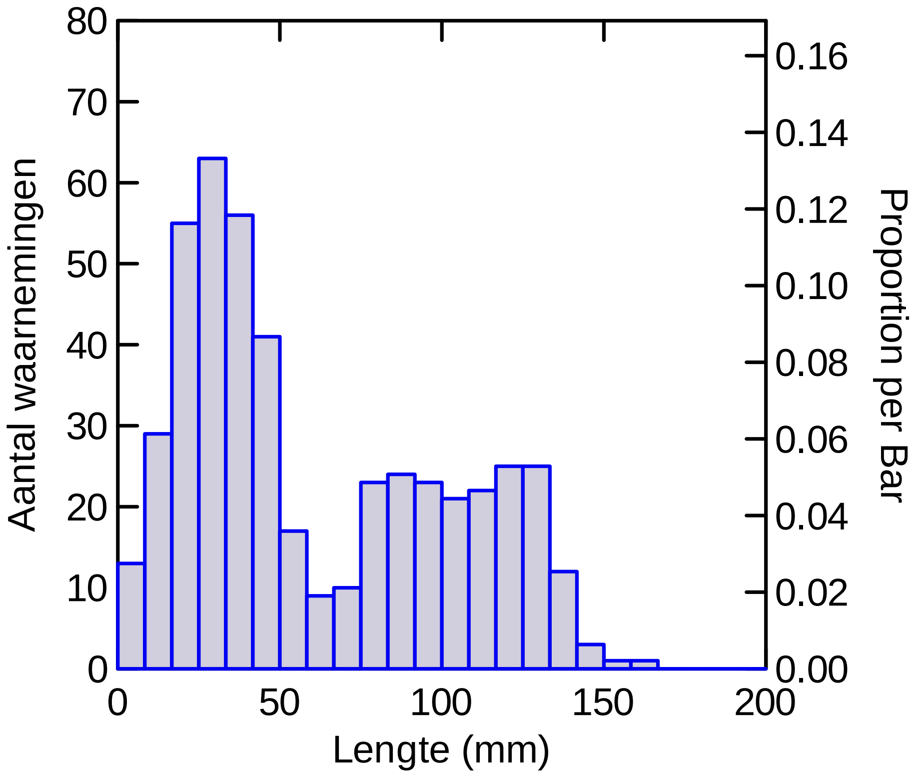
4.3.2 Total Biomass

Evaluation criterion	Evaluation result	Score
Indicator design	Total biomass is an important indicator in food chain models. There are various ways in which NEA GIG countries have interpreted the demand for an abundance indicator in the normative definitions. In the BEQI, total biomass per m^2 is calculated and assessed (Van Hoeij 2007, Twisk 2009).	5
WFD compliance	Yes, abundance indicator	5
Practical aspects	Biomass data are often estimated using length-biomass distributions, which makes the determination of biomass data less straightforward than total density data. Not all NEA GIG countries collect biomass data.	3
Sample area sensitivity	Clearly, total biomass is very dependent on the sample area which is a disadvantage.	3
Indicator calibration	Has been calibrated. The calibration of total biomass is a problem. First, total biomass is especially related to the eutrophication status (Dekker, 1984). Second, according to the Pearson Rosenberg model, lower total biomass may equally well indicate lower human pressures as a heavily polluted habitat (two sided distribution). Finally, large seasonal variations of biomass of species may occur (Essink et al. 1986). These phenomena makes it difficult to calibrate total biomass compared with a natural reference on a historical time scale, which is necessary for the WFD. Similar statements about the problems with calibration of total biomass have been made by Borja et al. (2000) and Meyer (2007).	3
Human pressure calibration	Biomass changes in Dutch coastal waters have been related to the invasion on *Ensis directus*. Clearly other pressures can affect total biomass. It may be difficult to discriminate human and natural pressures.	3
EQR normalization	Once the indicator has been calibrated, EQR normalization is possible.	3
Intercalibration	The BEQI is the only metric that uses biomass. This has led to an impaired intercalibration of BEQI for coastal waters in the past (EU, 2007)	2
Conclusions and average score	The calibration problem of the indicator total biomass and the intercalibration problem prevents its applicability as an independent WFD indicator.	3.4

4.3.3 Length distribution

Evaluation criterion	Evaluation result	Score
Indicator design	Size distributions give relevant information on the population structure of a species, usually the larger species such as bivalves, sea stars, crabs, etc. For example, years with high and low recruitment can be distinguished, and pressures on bivalves may reduce the fraction of the larger size classes. Furthermore the length of a bivalve determines it availability as food for see ducks. Size spectra are frequently reported for bivalves, see Figure X below. However, this indicator is only applicable to a part of the benthic organisms.	3
WFD compliance	No. Length is not suitable as a diversity, density or sensitive species indicator for the WFD because its information is too specific.	1
Practical aspects	Length information of bivalves is regularly available in Dutch monitoring programmes. The transformation of length information to age information requires a calibration per species. The use of an age parameter results in better understandable information.	3
Sample area sensitivity	This indicator is sensitive to the sample area, because sufficient individuals have to be collected to obtain a clear length distribution.	3
Indicator calibration	Has partly been calibrated to trawling disturbance (Jennings et al. 2001). Calibration is possible, but size calibration may not always be straightforward.	3
Human pressure calibration	Size distributions can give information on human pressures and recruitment processes. Smaller sizes may indicate a more pressured system, e.g. due to fisheries, or in case of shellfish due to recruitment.	3
EQR normalization	EQR normalization of length distributions does not seem straightforward.	2
Intercalibration	Since length distributions are not used in the NEA GIG metrics, intercalibration is expected to be difficult.	1
Conclusions and average score	Not useful as an independent WFD indicator. However, for other ecological monitoring and assessments projects, such as the Voordelta project of RWS and the Marine strategy, it gives important information on population structure, dynamics and human pressures.	2.4

Figure 4.4: Average length distribution from Ensis directus in the Voordelta Dutch coastal area in 2009, from Craeymeersch et al. (2010).



This figure shows the specimens < 1 year (<70 mm) and the specimens of 2 or more years old (>70 mm). This indicator gives information on the population structure, and on the availability of specific length classes as suitable food source for sea ducks.

4.3.4 Age distribution

Evaluation criterion	Evaluation result	Score
Indicator design	Age distribution of individual species (especially bivalves). These can be obtained by counting year rings on the shells of bivalves. Age information can be understood and interpreted well. This indicator is limited to species of which the age can be determined, usually bivalves. This is a useful indicator to indicate pressures on a species. Pressures such as fisheries usually decrease the age distribution. The application of this indicator is very limited, since age is commonly only measurable in bivalves.	2
WFD compliance	No, because this indicator is to limited to only bivalves.	1
Practical aspects	The reliable determination of year rings may be difficult.	3
Sample area sensitivity	This indicator is sensitive to the sample area, because sufficient individuals have to be collected to obtain a clear length distribution.	3
Indicator calibration	Has not been calibrated. In littoral transitional waters these year rings are often clearly visible. However, in sublittoral coastal waters growth often is a more steady process, and year rings may be	2

Evaluation criterion	Evaluation result	Score
	difficult to distinguish and data may be less reliable. Therefore there may be complications in the routine monitoring of age via year rings, and monitoring length distributions is recommended as a more straightforward, objective and reliable method.	
Human pressure calibration	Lower age distributions may indicate physical disturbances such as fisheries.	3
EQR normalization	This could be possible but is not straightforward	2
Intercalibration	This would be difficult, since this indicator is not used in NEA GIG metrics.	1
Conclusions and average score	Not useful as an independent WFD indicator.	2.1

4.3.5 Abundance Biomass Comparison (ABC) and Ratio (ABR)

Evaluation criterion	Evaluation result	Score
Indicator design	k-Dominance plots for a species based on biomass and abundance are plotted in the same figure (x-axis is log scale). Based on theoretical considerations, the relative position of the biomass curve will become lower than the density curve due to pollution and/or disturbance (Warwick 1994). A possible simplification of this model is to use the ratio of the total biomass and total abundance. Furthermore, this curve could be quantified into a single number by taken the number of species at 50% cumulative abundance in the ABC-curve. The concept is ecologically interesting and relevant, but Beukema et al. (1988) concluded that an ecotope-specific assessment of these curves is necessary.	5
WFD compliance	These indicators could be used as an abundance indicator if numerical values are used.	5
Practical aspects	k-dominance plots are not very easy to handle. ABR can be calculated very easily. Biomass data are not commonly measured	ABC: 2 ABR: 2
Sample area sensitivity	It is expected that these parameters are sensitive to the sampling area.	3
Indicator calibration	Clearly the ABC curves are difficult to handle numerically. Therefore, the use of a ratio of total biomass and total density (gram per specimen of a species), or the use of the number of species at 50% total abundance is recommended.	ABC: 2 ABR: 4
Human pressure calibration	This indicator is in principle able to detect human pressures. However the response of this indicator to human pressures may not be very clear.	3
EQR normalization	Possibly good, especially using the ABR	ABC: 3 ABR: 4
Intercalibration	The ABC is not used by the NEA GIG countries. This could lead to an impaired intercalibration.	ABC: 2 ABR:

Evaluation criterion	Evaluation result	Score
		2
Conclusions and average score.	This interesting indicator could be WFD compliant, but is not directly applicable. However, it is recommended to test the possible use of the indicator Abundance Biomass Ratio in future projects.	ABC: 3.1 ABR: 3.5

4.4 Sensitive/opportunistic species indicators

Sensitive/opportunistic species indicators are an essential part of the WFD metric for macrozoobenthos (see the normative definitions), because a major aspect of the WFD is to decrease the amount of human pressures and restoring good conditions. The AMBI is a popular index to calculate the degree of disturbance based on the species composition. Larger proportions of tolerant and opportunistic taxa indicate a disturbed benthic community. However, other indexes for disturbance have been developed or proposed, such as the BQI, the BOPA and the Bray-Curtis similarity index. A relatively new sensitive/tolerant species index which is currently investigated by Germany and Belgium is the AeTV.

Below, the NEA GIG disturbance indicators, and several other interesting options, are evaluated.

4.4.1 AMBI

Evaluation criterion	Evaluation result	Score
Indicator design	AMBI. For each species, a sensitivity/tolerance class (I to V) is assigned (Borja et al. 2000). I refers to species which are sensitive to organic matter; II less sensitive species, III tolerant species, IV second order opportunistic species which benefit from the organic matter and V first order opportunistic species. The numerical model is based on the Pearson Rosenberg model, intuitively logical and straightforward. Using the relative densities of the species and their sensitivity class scores, an integrated biotic index is calculated which reflects the relative amounts of sensitive, tolerant and opportunistic species in the benthic community. It is possible to calculate AMBI values based on biomass data (Muxika et al. 2010). These values correlate reasonably well with AMBI values based on abundance data.	4
WFD compliance	Yes, because the fractions of sensitive, tolerant and opportunistic species are calculated.	5
Practical aspects	AMBI scores can be calculated straightforwardly using user friendly and freely available software. In order to improve the AMBI species classifications and add two additional AMBI's designed for fisheries and sedimentation, a significant calibration effort must be performed.	5
Sample area sensitivity	The AMBI scores appear to be little affected by sample area (Fleischer et al. 2007) which is an important advantage. However, it is important to note that the AMBI puts the following minimum demands on the input data in order to generate an AMBI value with sufficient reliability: (a) at least 80% of the density has to be	5

Evaluation criterion	Evaluation result	Score
	assigned to AMBI classes, (b) the number of species has to be at least 3 and (c) the number of individuals has to be at least 6 (AZTI, 2010). It has been reported, probably due to these minimum demands, that the AMBI does not work well in mobile sand ecotopes, which have poor communities. This problem appears to be solved when data pools of 0.19 m^2 are used.	
Indicator calibration	In Borja et al. (2011) it was reported that the use of class I (sensitive species) alone appeared more sensitive to a human pressure index (R^2 = 0.51) than the total AMBI (R^2 = 0.23). The AMBI group approach is recommended for further testing.	4
Human pressure calibration	The AMBI appears to be especially useful for the pressures organic material and associated micropollutants. However, using species classifications for the pressures fisheries and sedimentation, two additional pressure selective AMBI's are currently developed for RWS Waterdienst (Gittenberger, 2011). Recently, Borja et al. (2011) found that the overall correlation of AMBI alone with a pressure index gradient is not very good (R^2 = 0.23). In this study, it was also reported that the use of AMBI group I correlated much better with the same pressure gradient (R^2 = 0.51). In Borja et al. (2008) a principal component plot shows, that for the dataset studied AMBI groups II and III correlate strongly with the Shannon index, while group I and V are the least correlated with Shannon. Therefore, group I and V seem to give the highest added value in a metric combined with the Shannon index. This leads to the conclusion that the use of AMBI groups may give more selective and discriminative information that the total AMBI. The indicator appears to be mostly calibrated to the human pressures organic material and associated micropollutants. The available species classifications (http://ambi.azti.es/) appear to be based on expert judgement, are not documented with publicly available scientific literature and are in some cases disputed by Dutch benthic experts. However, modifications of these classifications may be possible on request to Borja. A validation of these species classifications before application in the Dutch water bodies is necessary. This validation is performed in a current RWS/Gimaris project.	3
EQR normalization	A very useful aspect of the AMBI is that it has a clear WFD like classification in five classes (Salas et al. 2004). An AMBI score can be converted simply into an EQR score as follows: EQR = (7 – AMBI) / 7. This AMBI score must in principle be divided by a reference AMBI score. For example, without this reference correction a few ecotopes in transitional water would probably score only moderate in the reference state. This univariate calibration model is used in the NQI, DKI and IQI (Josefson 2009). The AMBI-classification can be used easily to calculate EQR-values and WFD classes in monovariate (e.g. the NQI, DQI and IQI, Josefson 2009) or multivariate (m-AMBI or BAT) linear regression models. Reference conditions can e.g. obtained by using the lowest AMBI score in a comparable data set (e.g. from the same salinity zone)	5

Evaluation criterion	Evaluation result	Score
	and add a safety margin of 10-15% (see e.g. Borja et al., 2008). Another option is to remove the opportunistic species (class IV and V) from a reference sample data set, and use its AMBI value as reference value.	
Intercalibration	Since many NEA-GIG countries use the AMBI, it can be intercalibrated well. With two added AMBI's for fisheries and sedimentation, the intercalibration could become slightly less, but is still expected to be good.	5
Conclusions and average score	The AMBI is in principle a valuable ecological indicator for use in the WFD. However, the species classifications must be validated with literature and documented expert judgement. The human pressure selectivity in Dutch marine waters can probably be improved by adding two pressure selective AMBI's for fisheries and extraction/sedimentation. Furthermore, the use of partial AMBIs may further improve the human pressure selectivity.	4.5

4.4.2 Bray-Curtis similarity index

Evaluation criterion	Evaluation result	Score
Indicator design	The Bray-Curtis similarity index is a well-established model to quantify the mathematical distance between the species composition of benthic samples. This index uses all species data and relative densities. In the BEQI the Bray-Curtis index is used as an indicator for sensitive, tolerant and opportunistic species (Van Hoey et al. 2007). A principle disadvantage of this model for the WFD is that it only quantifies the amount of change of a community, and does not tell if the detected change is an improvement or impairment.	4
WFD compliance	Probably no. This index does not indicate sensitive, tolerant or opportunistic species at the global level of the similarity index. However, if the Bray Curtis index is analyzed at a species level (see webtool on www. beqi.eu) than information at the species level can be obtained. However this information is not used to calculate the similarity index.	1
Practical aspects	Once a suitable set of reference data is available, this indicator can be calculated conveniently on the BEQI website (see www.beqi.eu).	4
Sample area sensitivity	This indicator is probably not very sensitive to sample area.	3
Indicator calibration	Has partly been calibrated. This index must be calibrated with a reference dataset. It can be difficult to find a suitable set of reference samples.	3
Human pressure calibration	This index shows variations probably due to human pressures.	3
EQR normalization	EQR normalization is not straightforward..	3
Intercalibration	Only the BEQI uses this index, and this index is not based on ecology but on statistical change. Therefore, an impaired intercalibration is expected	3
Conclusion and average score	This indicator is not suited as a sensitive/tolerant species indicator.	3

4.4.3 BOPA

Evaluation criterion	Evaluation result	Score
Indicator design	In the BOPA the ratio between the fraction of opportunistic polychaetes and sensitive amphipods is calculated (Dauvin & Ruellet 2007). The index can give very relevant ecological information for selected species groups. The species group concept can e.g. be applied to a ratio sensitive bivalves/opportunistic polychaetes.	5
WFD compliance	Yes, as a sensitive/tolerant/opportunistic species indicator.	5
Practical aspects	The model calibration requires careful work. After that, this type of indicator can be used straightforward.	4
Sample area sensitivity	This indicator may be relatively sensitive to sampling area due to species numbers.	3
Indicator calibration	A careful assignment of species to sensitivity classes is necessary. This is possible.	3
Human pressure calibration	The BOPA appears to give sensitive signals for the pressures organic matter, heavy metals and hydrocarbons. In Borja et al. (2011) a relatively good correlation of AMBI group I with a human pressure gradient ($R^2 = 0.51$) has been reported.	4
EQR normalization	This index can be harmonised well (comparable set up as for AMBI.	5
Intercalibration	Species groups are not officially used in the NEA GIG metrics at present. This could lead to a slightly lower intercalibration performance .	4
Conclusion and average score	The species group concept is important because it can give more sensitive signals and can correlate relatively well with human pressures if appropriate species groups are selected.	4.1

4.4.4 Estuary-type method

Evaluation criterion	Evaluation result	Score
Indicator design	This indicator (Krieg 2010) uses an open list of indicator species which are characteristic for specific estuaries, or frequently visit the estuary. Each species is given an eco-value from 1 to 5, where 5 indicates the strongest relation of a species with the estuary, i.e. genuine brackish water species and often more rare. Species with eco = 1 are very common and also occur in other water types.	5
WFD compliance	Yes; because sensitive, tolerant and opportunistic species are defined.	5
Practical aspects	The calibration process is labour intensive. This method uses a special sampling procedure with 0.25 mm sieving. This makes the sampling much more time-consuming.	2
Sample area sensitivity	The method presumably is not very sensitive to the sample area.	4
Indicator calibration	The indicator appears to have been calibrated carefully with a lot of (partly historical) ecological knowledge.	5
Human pressure calibration	This indicator probably is sensitive to human pressures. It seems that no direct link of the indicator response to specific human pressures can be made.	3
EQR normalization	Is well possible.	5

Evaluation criterion	Evaluation result	Score
Intercalibration	Germany is the only country which uses this method. However, comparable results of this indicator with the m-AMBI have been presented by Germany.	3
Conclusions and average score	This indicator appears to be of high ecological quality due to its careful calibration. The labour intensive 0.25 mm sieving however is a practical disadvantage. The link with human pressures does not appear to be very evident.	3.9

4.4.5 Infaunal trophic index

Evaluation criterion	Evaluation result	Score
Indicator design	This indicator characterizes the trophic structure of a benthic community (Word 1978, Lavaleye 1999). Species are classified in four classes: suspension feeders, interface feeders, deposit feeders and subsurface deposit feeders. Suspension feeders are seen as indicators of a good community status, while deposit and especially subsurface deposit feeders as indicators for an impaired status. The fractions of these four classes are integrated into a score (0-100). The model appears to be ecologically logical and meaningful, and has sufficient discrimination using four classes. The ITI is frequently reported in the literature (Pinto et al. 2009) and has been reported to give relevant ecological information.	5
WFD compliance	Yes; it is an indicator of sensitive and opportunistic species.	5
Practical aspects	The calibration may be time-consuming. Once this indicator has been calibrated, it can be calculated very easy.	4
Sample area sensitivity	It is expected that the ITI is not very sensitive to the sampling area.	4
Indicator calibration	Has been calibrated. Species have to be classified into four classes. In some cases, species may have multiple feeding routes, and classification may not be clear.	4
Human pressure calibration	The ITI gives varying results. Some studies show relations with pressures, others not. It is expected to react to the human pressures sedimentation (disturbance of filter feeders) and fisheries (damage of filter feeders). It could be that the ITI correlates well with AMBI-fisheries and AMBI-sedimentation.	3
EQR normalization	Linear two point calibration (EQR 1 and 0) is possible.	5
Intercalibration	The ITI is not used in the NEA GIG countries, but could be possible.	4
Conclusions and average score	The ITI is a relevant indicator which seems to be related to the human pressures sedimentation and fisheries.	4.3

4.4.6 R/K-strategists

Evaluation criterion	Evaluation result	Score
Indicator design	Classification of species according to their survival strategy as r-strategists (small, fast reproducing, many off-spring, and tolerant species) or K-strategists (large, slow reproduction, low number of off-spring, sensitive). The concept is basically very logical and ecologically relevant. This indicator has been reported in the literature frequently (Lavaleye 1999). The classification in only two	4

	groups limits the sensitivity of this indicator. E.g. the ITI uses four classes.	
WFD compliance	Yes; indicator for sensitive, tolerant and opportunistic species	5
Practical aspects	It may be difficult to assign species correctly. The use of this index is very simple.	3
Sample area sensitivity	This indicator is probably not very sensitive to the sampling area, since only two classes are used.	4
Indicator calibration	Has been calibrated. There are several possible criteria to classify species as r- or K-strategist. This poses a problem in the standardized classification of species as r- or K strategist, which is also observed in the literature.	4
Human pressure calibration	An increase in the fraction of r-strategists is interpreted as an increase in human pressures. The types of human pressures however cannot be distinguished. In the GONZ studies (Kabuta 2000) the r/K-ratio results were not as good as the ITI.	2
EQR normalization	This indicator can be calibrated using linear two point calibration.	5
Intercalibration	This indicator is not used in the NEA GIG metrics, but probably has a good intercalibration potential.	4
Conclusions and average score	The r/K-strategy is a relevant indicator which could be of some use for the WFD.	3.9

In chapter 5, an overview and conclusion is given of the evaluation results for the indicators.

4.5 Evaluation of WFD metrics for marine benthos

Indicators are in principle univariate. In order to meet the WFD normative definitions, at least three indicators have to be determined (in principle for diversity, abundance and sensitive/tolerant species). An important question is, how these indicators can be integrated into a metric in order to get the most sensitive and discriminative EQR results for the quality of benthic communities in relation to human pressures. The available indicator WFD metrics are described and evaluated using the following criteria:

- Metric design. The indicator composition, integration and ecological information of the metric are evaluated.
- WFD compliance. Is the metric WFD compliant?
- Practical aspects. How easy can the metric be used? How large are the data and calculation demands?
- Sample area sensitivity. How sensitive is the metric to the sampling surface area?
- Metric calibration. How well can the metric be calibrated?
- Human pressure calibration (sensitivity and correlation). Has the metric been calibrated with a human pressure index? How sensitive is the response of the metric to the pressure index? What is the correlation coefficient?
- EQR normalization. How well can the metric response be calibrated into an EQR value?
- Intercalibration. How well can the metric be intercalibrated?
- Conclusions about the metric and average score

4.5.1 BEQI

Evaluation criterion	Evaluation result	Score
Metric design	The BEQI has been described in the Dutch metric document (Van Hoeij et al. 2007). The BEQI is composed of three levels (see chapter 1): • level 1. A first problem of level 1 metric is that at the moment practically there no data available for primary production due to lack of a reliable monitoring method; currently only estimated values are used with unknown reliability. In addition, the information value and need for this indicator is questioned by Prins (Theo Prins, pers. comm.), since a phytoplankton/benthic system will in principle always reach the specified equilibrium ratio, although in terms of 1 to 3 years this relation may be unbalanced. Finally, this indicator seems to be superfluous since phytoplankton is already assessed for the WFD in a separate metric, and double assessments are discouraged in the WFD. • level 2. In the present BEQI, absolute areas of several habitats (littoral, shallow water, sand banks, mud banks, littoral mussel banks) are compared to reference areas of these habitats from around the year 1900. It appears that the importance and correctness of this level 2 submetric is undisputed, both in the Netherlands and in the EU. It is an option to add the habitat sublittoral mussel banks, and possibly *Spisula subtruncata*, to this metric level. • level 3. The intercalibration is only focused on level 3 of the BEQI. As evaluated in the indicator sections, the abundance indicators total density and total biomass have a lower score and are not very suitable for the WFD metric. Also, the Bray Curtis index is not suited as an indicator for sensitive/opportunistic species. Species richness is a correct and useful indicator.	2
WFD compliance	No. The Bray Curtis index is not suited as an indicator of sensitive and opportunistic species.	1
Practical aspects	There is a useful webtool available at www.beqi.eu. The comprehensiveness of this tool for non-specialists is limited.	3
Sample area sensitivity	The indicator density, biomass and species richness are sensitive to the sampling area.	3
Metric calibration	Has been calibrated. The BEQI requires extensive reference data sets. Furthermore, the statistical calculation procedure of reference data is complex and not very transparent. The availability of a web-based calculation procedure has partly mitigated this problem (see www.beqi.eu).	3
Human pressure calibration	The Bray Curtis index indicates objective change of a community, but does not specify its correlation with community quality and human pressures. A recent study of Borja et al. (2011) shows that the BEQI correlates at best moderately with a human pressure index ($R^2 = 0.24$).	3
EQR normalization	The selection of a reference data set with a sufficient number of sample pools is a problem, since suitable reference data are rare. The statistical approach to set e.g. the 2.5 or 5 percentile as the	3

Evaluation criterion	Evaluation result	Score
	good/moderate boundary can be disputed ecologically (see www.beqi.eu). The use of linear calibration with a suitable single point reference sample appears to be more logical and straightforward.	
Intercalibration	In the intercalibration of the BEQI for coastal waters it appeared to be one of the metrics producing the most deviant scores. For the intercalibration of the BEQI in transitional waters, the same problem is expected.	2
Conclusions and average score	The BEQI at level 3 shows several problems. The use of the Bray Curtis index is not WFD compliant.	2.5

4.5.2 M-AMBI

Evaluation criterion	Evaluation result	Score
Metric design	The m-AMBI metric is based on the indicators species richness S, Shannon index H' (log base 2) and the AMBI (Muxika et al. 2007). It has been shown in Borja et al. (2011) that S does not correlate as good with human pressure index (R^2 = 0.38) as H' (R^2 = 0.64). Furthermore, it can be argued that it would be more balanced if a single diversity indicator and a single sensitive species indicator are combined, instead of two diversity indicators and a single sensitive species indicator. Therefore, it is possible that a metric with only H' and AMBI can be more balanced and efficient.	3
WFD compliance	Yes. The WFD demand for an abundance indicator has been interpreted as relative species densities in both H' and AMBI.	5
Practical aspects	User friendly software is freely available from the AMBI website. The use of m-AMBI is easy. The use of user-customized species sensitivity classes in a coming update will further improve the user friendliness.	5
Sample area sensitivity	The indicators AMBI and H' are little sensitive to the sampling surface. S is very sensitive for the sampling surface.	3
Metric calibration	Has been calibrated. The AMBI species sensitivity classes need to be validated. The model uses a two point multivariate calibration, with a reference point (EQR = 1) and a bad point (EQR = 0). Sample points in the three dimensional space are projected perpendicular on this calibration line, and the EQR-value is calculated using an equidistant EQR scale. This is a sophisticated calibration method.	3
Human pressure calibration	The sensitivity of AMBI itself to human pressures appeared to be not good in a recent study (Borja et al. 2011). This is probably caused by the fact that the AMBI is not calibrated to physical pressures such as fisheries and sedimentation, but mostly to the pressure organic material and micropollutants. Also, the use of AMBI group I appears to give a better pressure correlation. However, in combination with H' and S a reasonably good human pressure sensitivity was reported for the m-AMBI (R^2 = 0.60), but only for a subset of human pressures, not fisheries or extraction/sedimentation.	3
EQR normalization	The model uses a two point multivariate calibration, with a	5

Evaluation criterion	Evaluation result	Score
	reference point (EQR = 1) and a bad point (EQR = 0). Sample points in the three dimensional space are projected perpendicular on this calibration line, and the EQR-value is calculated using an equidistant EQR scale. It has been shown however that the NQI, which is a linear combination of two univariate indicators H' and AMBI, gives comparable human pressure sensitivities with the m-AMBI and BAT (Borja et al. 2011). The calculation procedure for the NQI, DKI and IQI is very simple, which may be an advantage over the multivariate m-AMBI procedure for easy automation, script development etc.	
Intercalibration	The intercalibration potential of the m-AMBI is probably good.	5
Conclusions and average score	The m-AMBI appears to be a fairly good metric, but it can possibly be improved by omitting the indicator S, by using additional pressure selective AMBIs.	4

4.5.3 NQI

Evaluation criterion	Evaluation result	Score
Metric design	The Norwegian Quality Index is calculated as follows: NQI = 0.5 * (1 – AMBI/7) + (0.5 * (SN/2.7) * (Ntot/Ntot + 5)) in which SN is ln(S)/ln(ln(N)) and Ntot is the number of individuals in the sample. The SN is a species richness indicator which is corrected for the number of individuals. In this way it becomes less sensitive to the sample area. 2.7 is the maximum (reference) value of SN observed in the samples. In Borja et al (2011) it was reported that SN correlates reasonably well with a pressure index (R^2 = 0.565), better than S (R^2 = 0.376), but not as good as H' (R^2 = 0.637). Therefore, it can be expected that this metric can be improved by using H' instead of SN. It is remarkable in this formula that a reference value for AMBI is not used, although in AMBI calibration this is commonly needed. The last part of the formula is a correction factor for a low number of individuals in the sample. In this metric the AMBI and diversity indicator have an equal weight, which can be considered to give a balanced evaluation using two different types of indicators. The metric is also relatively simple.	3
WFD compliance	Yes, because it contains indicators for diversity, abundance and sensitive species.	5
Practical aspects	Calculations are relatively simple and can be automated easily.	5
Sample area sensitivity	The AMBI is not sensitive to sampling area. SN is less sensitive than S for sampling area.	4
Metric calibration	Has been calibrated. The AMBI has to be calibrated using species classifications, which may not yet be optimal. Pressures such as fisheries and sedimentation may not be adequately assessed.	3
Human pressure calibration	A reasonably good human pressure correlation (R^2 = 0.59) has been reported (Borja et al. 2011).	4
EQR normalization	This is simple, because the AMBI and SN are calibrated using univariate two-point linear calibration.	5
Intercalibration	This was good in the coastal water calibration process. In Borja	5

Evaluation criterion	Evaluation result	Score
	2007 a good comparability of the NQI with the DKI, UK-metric and m-AMBI has been demonstrated.	
Conclusions and average score	The NQI appears to be a simple and reasonably human pressure sensitive metric. It can probably be improved by using H' instead of SN, like in the DKI and IQI. The AMBI calibration can be improved.	4.3

4.5.4 DKI

Evaluation criterion	Evaluation result	Score
Metric model	The DKI is calculated as follows: DKI = (((1- AMBI/7) + (H'/Hmax)) / 2) * ((1 – 1/N) + (1 – 1/S))/2 This formula is quite similar to the NQI, but instead of SN the Shannon index is used which probably is the best choice for a diversity indicator (Borja et al. 2011), and the correction factor is applied both for the number of species and individuals. Using a standardized sample area of 0.19 m^2, the correction factor is probably small due to sufficiently high values of N and S. It is remarkable in this formula that a reference value for AMBI is not used, although in AMBI calibration this is commonly needed.. This may be incorrect.	3
WFD compliance	Yes, because it contains indicators for diversity, abundance and sensitive species.	5
Practical aspects	Calculations are relatively simple and can be automated easily.	5
Sample area sensitivity	The AMBI is not sensitive to sampling area. H' is less sensitive than S for sampling area.	4
Metric calibration	Has partly been calibrated. The AMBI species sensitivity classifications have to be validated and extended to the pressures fisheries and sedimentation.	3
Human pressure calibration	The human pressure sensitivity has not been validated in the study of Borja et al. (2011), but is expected to be higher than of the NQI due to the use of H' instead of SN, where H’ performs better than SN on this criterion.	4
EQR normalization	This is simple, because the AMBI and H' are calibrated using univariate linear calibration.	5
Intercalibration	The DKI showed the best intercalibration results in a study of Borja et al. (2007) which compared the m-AMBI, the NQI, the DKI and the UK-metric.	5
Conclusions and average score	The DKI is very similar to the NQI, and could perform even slightly better than the NQI due to the use of H'.	4.3

4.5.5 IQI

Evaluation criterion	Evaluation result	Score
Metric design	The IQI is calculated as follows:	3

	$IQI\,v.IV = \left(\left(0.38\times\left(\frac{1-(AMBI/7)}{1-(AMBI_{REF}/7)}\right)\right)+\left(0.08\times\left(\frac{1-Lambda'}{1-Lambda'_{REF}}\right)\right)+\left(0.54\times\left(\frac{S}{S_{REF}}\right)^{0.1}\right)-0.4\right)\Big/0.6$ The indicators in the IQI have been selected via a principal component analysis, which shows that AMBI, S and the diversity parameters are three clearly distinguishable indicator groups. The formula has been optimized in a statistical evaluation study (Phillips pers. comm.). A difference with the NQI and DKI is the use of the Simpson index, which did not give a good correlation with a human pressure index (R^2 = 0.26, Borja et al. 2011). Furthermore, a smaller weight factor is used for λ which is a significant difference. In view of the good correlation of H' with a human pressure index (Borja et al. 2011), it seems better to use at least equal weight coefficients. An important difference with the DKI and NQI is, that a reference value for AMBI is used. This is probably essential, since for example in transitional waters many tolerant (class III) species occur naturally, which could probably lead to a moderate classification in a few ecotopes in the natural state.	
WFD compliance	Yes, because it contains indicators for diversity, abundance and sensitive species.	5
Practical aspects	Calculations are relatively simple and can be automated easily.	5
Sample area sensitivity	The AMBI is not sensitive to sampling area. λ is less sensitive than S for sampling area. However, S in the metric is sensitive to the sampling area.	3
Metric calibration	Has party been calibrated. The AMBI has to be calibrated using species classifications, which may not yet be optimal. λ does not have to be calibrated.	3
Human pressure calibration	The human pressure sensitivity of the UKI has not been validated in the study of Borja et al. (2011), but could be somewhat lower due to the use of lambda as compared to H' (Borja et al. 2011)	3
EQR normalization	This is simple, because the AMBI and λ are calibrated using univariate linear two-point calibration.	5
Intercalibration	The UK-metric showed the good intercalibration results in a study of Borja et al. (2007) which compared the m-AMBI, the NQI, the DKI and the UK-metric.	5
Conclusions and average score	The UK-metric is more or less comparable to the NQI and DKI, but could perform slightly less than the NQI due to the use of λ, and a lower weight factor for λ.	4

4.5.6 BQI

Evaluation criterion	**Evaluation result**	**Score**
Metric concept	The BQI is calculated as follows: BQI = Σ (Ni/Ntotal * Sens.value i) * (log10 (S+1)) * (Ntotal/(Ntotal + 5)) The BQI is based on the $ES50_{0.05}$ (expected number of species at 50 individuals) sensitivity values of species (Rosenberg et al. 2004). The rationale behind this model is that the sensitivity of species is reflected by the number of species which are found in the sampled ecotope. High species numbers indicate a relatively undisturbed ecotope, which contains relatively many sensitive	4

	species. In contrast, ecotopes with low numbers of species apparently are disturbed and contain mainly tolerant species. The $ES50_{0.05}$ model is intuitively logical and appears to be scientifically undisputed. A limitation of this model may be that the evenness of the community is not accounted for, because it is purely based on the number of species.	
WFD compliance	Yes. The BQI contains an indicator for species sensitivity, diversity and abundance (as a correction factor for low numbers of individuals).	5
Practical aspects	The calculation of ES50 values is data and labour intensive. Furthermore, these values have to be re-calculated for different marine regions.	3
Sample area sensitivity	The BQI is based on a standardized sample area of 0.1 m^2.	5
Metric calibration	Has been calibrated. Species sensitivity values can be calculated using an objective scientific method which is an advantage.	5
Human pressure calibration	The model has been developed in Swedish fjords, but will in principle also work for other types of pressures. A disadvantage of the BQI seems that human and natural pressures cannot be discriminated and that the relationship with other pressures than eutrophication/oxygen is unclear. For example, species which live in intertidal regions with heavy natural pressures also have low sensitivity values. This is a drawback for the WFD. The correlation of the BQI with a pressure index gradient was reasonably good (R^2 = 0.42), but was not among the highest correlating indicators and metrics (Borja et al. 2011).	3
EQR normalization	The EQR normalization uses the good-moderate boundary obtained from pressure-impact plots. This is an interesting and probably valid approach.	5
Intercalibration	Only Sweden uses the BQI. The BQI has been compared with the AMBI and appears to give often different results (Grémare et al. 2009). It is estimated from several results that the BQI can be intercalibrated, but some systematic differences from AMBI-based metrics may occur.	3
Conclusions	The BQI is an objective tool to assess the sensitivity of species. However, the sensitivity of individual species is not clearly related to human or natural pressures. Furthermore, the BQI only focuses on the species richness, and does not take evenness into account.	4.1

4.5.7 BAT

Evaluation criterion	**Evaluation result**	**Score**
Metric design	The BAT metric is based on the indicators Margalef's d, Shannon index (log base 2) and the AMBI (Muxika et al. 2007). It has been shown in Borja et al. (2011) that d correlates somewhat better with a human pressure index (R^2 = 0.48) that S (R^2 = 0.38), which could make this metric slightly better than the m-AMBI. Furthermore, it is possible that it would be more balanced if a single diversity indicator and a single sensitive species indicator are combined, instead of two diversity indicators and a single sensitive species indicator. Therefore, it can be argued that a metric with only H' and	4

Evaluation criterion	Evaluation result	Score
	AMBI can be more balanced and efficient.	
WFD compliance	Yes. The WFD demand for an abundance indicator has been interpreted as relative species densities in both H' and AMBI.	5
Practical aspects	User friendly software is freely available from the AMBI website. The use of AMBI is easy. The use of user-customized species sensitivity classes in a coming update will further improve the user friendliness.	5
Sample area sensitivity	The indicators AMBI and H' are little sensitive to the sampling surface. D is less sensitive to the sampling area than S, which could be a small advantage of the BAT over the m-AMBI.	4
Metric calibration	Has party been calibrated. The AMBI species sensitivity classes need to be validated. Furthermore, the pressures fisheries and sedimentation may not be adequately assessed. The model uses a two point multivariate calibration, with a reference point (EQR = 1) and a bad point (EQR = 0). Sample points in the three dimensional space are projected perpendicular on this calibration line, and the EQR-value is calculated using an equidistant EQR scale.	3
Human pressure calibration	It has been reported in Borja et al. (2011) that the BAT performs equally well with respect to pressure sensitivity and correlation as the m-AMBI.	4
EQR normalisation	The model uses a two point multivariate calibration, with a reference point (EQR = 1) and a bad point (EQR = 0). Sample points in the three dimensional space are projected perpendicular on this calibration line, and the EQR-value is calculated using an equidistant EQR scale. It has been shown however that the NQI, which is a linear combination of two univariate indicators H' and AMBI, gives comparable human pressure sensitivities with the m-AMBI and BAT (Borja et al. 2011). The calculation procedure for the NQI, DKI and IQI is very simple, which may be an advantage over the multivariate m-AMBI procedure for easy automation, script development etc.	4
Intercalibration	The intercalibration potential of the m-AMBI is probably good.	5
Conclusions and average score	The BAT appears to be a fairly good metric, but it can possibly be improved by omitting the indicator S, by using additional pressure selective AMBI's.	4.3

In the next chapter, an overview and conclusion is given of the evaluation results for the metrics.

5 Conclusions and recommendations

From early steps in the WFD intercalibration process, it became clear that the benthic metric developed in the Netherlands and Belgium, the BEQI, was considerably different from the metric developed in other EU countries. Moreover, it did not fulfil the WFD criteria for metrics. As a result, there was a need in the intercalibration process of the Dutch metric for marine macrozoobenthos to evaluate this metric, and if possible to improve it. For this purpose, the Waterdienst Rijkswaterstaat asked Deltares, and in a later stage Gimaris, to review (a) the current BEQI, (b) other indicators and metrics which are used in the West European countries for the WFD assessment of marine benthos and (c) make a selection of other promising indicators and metrics. There were several questions about the suitability of the currently used indicators and metric at level 3 of the BEQI (especially the indicators total density, total biomass and Bray-Curtis index), and about the use of the level 1 assessment (ratio primary production / total benthic biomass).

Below are given the main conclusions and recommendations from this review study with some discussion about the evaluation process and its results. Paragraph 5.1 concerns the single indicators. Paragraph 5.2 involves the metrics composed of various single indicators mentioned in part A. In paragraph 5.3, additional recommendations are given on sampling, analysis and assessment. The conclusions and recommended are numbered in order of appearance.

5.1 Indicator selection

In table 5.1 below, an overview is given of the evaluated indicators and their scores (on a scale of 1 to 5). The scores per criterion are the average of the individual scores of the three authors of this report. The average score is the unweighted average of all criteria scores. This may seem inappropriate since some criteria could (or should) be given more weight than others. One such criterion is the human pressure calibration. Without an indicator actually doing what it is supposed to –indicating pressures- any other score is redundant. We decided to leave the ranking on basis of the average score as is, but to add a red colouring to those indicators that score below 3 (poor) on pressure calibration. These poor scorings are mostly based a poor performance, not on lack of knowledge. This means that they are not likely to be proper indicators in the context of this document, unless additional studies clearly show otherwise. Nevertheless, it cannot be sufficiently repeated that even the highest scoring indicators on the pressure criterion still lacks proper calibration for the most important ‘Dutch’ pressures: bottom fishing and extraction/sedimentation. In order to tackle this problem, currently a project commissioned by RWS Waterdienst is finished which defines two new AMBIs, designed selectively to detect the human pressures bottom trawl fisheries and sedimentation, respectively (Gittenberger, 2011). These two indicator will be tested in the current calibration/intercalibration project, and in other RWS projects.

In the three following paragraphs, conclusions and recommendations are given on diversity indicators, abundance indicators and sensitive/tolerant/opportunistic species indicators, the three categories for which the WFD normative definitions demands indicators in the WFD metric.

Table 5.1: Evaluation of indicators which are potentially useful for the WFD. Red marked are the indicators that score poorly on human pressure calibration.

Indicator	Indicator design	WFD compliance	Practical aspects	Sample area sensitivity	Indicator calibration	Human pressure calibration	EQR normalization	Inter-calibration	Average score
Diversity indicators									
Shannon Wiener	5	5	5	4	5	4	5	5	4.8
Simpson	4	5	5	4	5	2	5	3	4.3
Margaleff's d	3	5	4	4	5	3	5	4	4.1
S - standardised area	3	5	3	3	5	3	5	5	4.0
S - ES50	3	5	3	4	4	3	4	4	3.8
Evenness	3	5	5	3	5	2	5	2	3.8
S - collectors curve	3	5	2	2	5	3	5	4	3.6
k-dominance	4	5	3	3	2	2	4	3	3.3
Abundance, size, age indicators									
Total density	5	5	5	3	3	3	3	3	3.8
Abundance-Biomass Ratio	5	5	2	3	4	3	4	2	3.5
Total biomass	5	5	3	3	3	3	3	2	3.4
Abundance-Biomass Comparison	5	5	2	3	2	3	3	2	3.1
Length distribution	3	1	3	3	3	3	2	1	2.4
Age distribution	2	1	3	3	2	3	2	1	2.1
Sensitive/opportunistic species									
AMBI standard	4	5	5	5	4	3	5	5	4.5
Infaunal trophic index	5	5	4	4	4	3	5	4	4.3
BOPA	5	5	4	3	3	4	5	4	4.1
AeTi	5	5	2	4	5	3	5	3	4.0
r/K strategists	4	5	3	4	4	2	5	4	3.9
Bray Curtis	4	1	4	3	3	3	3	3	3.0

1. Evaluation process. The individual scores of the three authors appeared to be quite comparable. In many cases the scores were identical, and in a limited number of cases a variation of 1 point from the average was observed. These differences mainly occurred for the review criteria: indicator design, indicator calibration and human pressure calibration. This evaluation comparability gives confidence in the conclusions which have been drawn in this review from the two evaluation tables for indicators and metrics. Nevertheless, the scores are for intercomparison only, and largely based on expert judgement. Thus, the evaluation in this report should be viewed as indicative and approximate.

2. Diversity indicators. The Shannon index with log base 2 clearly appears to be the best diversity indicator and is selected for use in the updated BEQI. The Shannon index is the most suitable diversity indicator in view of its relatively high human pressure sensitivity and correlation and its good indicator concept. This index value must be compared with a reference Shannon value for the assessed ecotope, being the highest value for H' which has been found in that ecotope (salinity zone, tidal/subtidal, sandy or muddy). But it should be noted that even the best scoring indicator (Shannon-Wiener) still lacks in calibration with the pressures assumed most important in Dutch transitional, coastal and offshore waters: bottom fisheries, sand extraction and coastal nourishment (sedimentation). In a follow-up study, this calibration will be done for samples from Dutch traditional and coastal waters.

It is also recommended to routinely co-calculate S and E as supporting information, not for WFD assessment, but to interpret changes in H'.

3. Abundance indicators. Apart from total abundance (score 3.8), the other abundance indicators have lower scores, especially due to their calibration problems. No other NEA GIG metric uses total abundance as an independent indicator. Instead, total abundance is used by several metrics (NKI, BKI, BQI) as a correction factor for undersampling. Furthermore, several other metrics such as the m-AMBI translate the WFD requirement for an abundance indicator into the use of relative abundances in the AMBI and Shannon index. In view of these findings, it is recommended not to use an independent abundance indicator in the BEQI update, but to use instead a density correction factor such as in the DKI and NQI. This abundance correction factor is a useful check and correction for a sufficient amount of sampling area, but may be omitted if it is small (e.g. <5%) in view of the use of a standardized sample pool area of 0.19 m^2. This will be checked in the coming (inter-)calibration project.

4. Sensitive/opportunistic species indicators. The standard AMBI appears to receive the highest review score. The BOPA, AeTI, and Infaunal Trophic Index also appear to receive good scores from 4.0 or higher.

The AMBI is in principle a logical and useful indicator for sensitive/opportunistic species. However, the standard AMBI must be calibrated better. Furthermore, the AMBI must be calibrated for more pressures, in specific the pressures fisheries and extraction/sedimentation which are very important in the Dutch marine water. In the coming data project these two additional AMBIs will be studied for their applicability to these human pressures in the Dutch assessment, and which of these is most suited for the Westerschelde intercalibration.

It was recently shown by Borja et al. (2011) that the standard AMBI appears not sensitive enough for an average human pressure index. In the past, it was also argued by Dutch benthic experts that the pressures fisheries and sedimentation are not well indicated by the standard AMBI. Therefore, in a current Dutch benthos project two additional AMBIs are being developed for the pressures bottom fisheries and sedimentation (extraction/nourishment), respectively. It is expected that these two additional AMBIs will give selective information on these two pressures which are common in Dutch coastal and transitional waters. The performance of these two AMBIs will be tested in a future calibration/intercalibration project in spring 2011.

Another option to increase the human pressure sensitivity of the AMBI is AMBI group I. This partial AMBI is clearly more sensitive for human pressures, and correlates better, than the standard AMBI (R^2 = 0.51; Borja et al. 2011, Borja et al. 2008). It must be noted however that the use of AMBI group I alone is not WFD compliant, since also opportunistic species have to be assessed for the WFD. It is an option to use the AMBI group approach, using at least the sensitive (group I and II) and opportunistic (III and IV) AMBI groups. This option will be tested in the coming (inter-) calibration project. However, the use of a complete AMBI probably remains preferable because a standardized and comparable setup of, and a complete, indicator is used.

5.2 Metric selection

In table 5.2 below, an overview is given of the evaluated WFD metrics and their scores.

Table 5.2: Evaluation of several WFD metrics. Average scores of the three authors are presented.

Metric	Metric design	WFD compliance	Practical aspects	Sample area sensitivity	Metric calibration	Human pressure calibration	EQR calibration	Inter-calibration	Average score
DKI	3	5	5	4	3	4	5	5	4.3
NQI	3	5	5	4	3	4	5	5	4.3
BAT	4	5	5	4	3	4	4	5	4.3
BQI	4	5	3	5	5	3	5	3	4.1
m-AMBI	3	5	5	3	3	3	5	5	4.0
IQI	3	5	5	3	3	3	5	5	4.0
BEQI	2	1	3	3	3	3	3	2	2.5

5. Conclusion on metrics. It appears from this table that a DKI/NQI-like metric is the most effective and efficient. Therefore this metric model, with a few modifications, is proposed in conclusion 11.

6. BEQI level 1. It is recommended to omit the BEQI level 1 assessment, since this assessment resembles a phytoplankton assessment which is already available as a separate, and more accurate, phytoplankton WFD metric. Therefore, the added value of this submetric is small. In addition, there are almost no monitoring data for primary production available and a lot of data assumptions have to be made at present.

7. BEQI level 2. The current BEQI level 2 assessment of the areas of littoral (sandy and muddy) habitat, shallow water and littoral mussel banks, is generally regarded as very valuable. Even a small extension at this level by adding the eco-element “area of sublittoral mussel banks” is proposed at this moment.

8. BEQI level 3. The current BEQI level 3 metric appears to score the lowest (average score 2.5) of all evaluated metrics (see Table above). This is mainly due to the fact that (a) the Bray Curtis index is not really an indicator for sensitive, tolerant and opportunistic species; (b) the

indicators total density and total biomass are difficult to calibrate and intercalibrate and (c) the indicator species richness is not as effective as the Shannon Wiener index.

9. Species richness indicator. It may not be necessary to add a species richness indicator to the proposed BEQI metric, because a very similar human pressure correlation is observed for H' (Borja et al. 2011). Furthermore, the NQI, which only contains AMBI and SN, gives a human pressures correlation very similar to the BAT and the m-AMBI. Also, a PCA plot from Borja et al. (2008) shows that H' and AMBI have opposite (very different) vectors and are therefore complementary indicators, while the S vector is not very different from the H' vector and therefore will not add much to a metric containing H' and AMBI. Furthermore, this choice largely eliminates the problem of the high dependence of S on the sampling area. On the other hand, it was shown in the Intercalibration report for Coastal water (EU, 2007) that the additional use of S in the m-AMBI and Margaleff's *d* in BAT led to a lower EQR scores than using the NKI and DKI which do not use S, and the m-AMBI and BAT scores are more in agreement with expert judgement on the benthic quality of Dutch coastal waters. The scientific status of this Intercalibration report is however unclear, and neither are the datasets and reference values which have been used. Therefore, the evaluation results for indicators and metrics from the recent pressure index paper of Borja et al. (2011) are considered leading. However, it remains a point of attention to validate the results of the new BEQI metric with expert judgement of a team of Dutch benthic experts. If necessary, the weight factors for H' and AMBI in the proposed new metric can be adjusted in order to reach a metric result which is in agreement with the Dutch expert judgement.

10.Total abundance/biomass ratio. It is recommended to test the indicator Total abundance/total biomass ratio (ABR) as a general human pressure indicator.

11. EQR-calibration. There are two principal options to calibrate a metric, namely linear combination of univariately calibrated indicators (as used in e.g. the NQI, DKI and IQI) and multivariate calibration (as used in the m-AMBI and BAT). Multivariate calibration appears to be in principle a scientifically correct method to use. However, linear combination of univariate indicators, which is somewhat less refined than multivariate calibration, appears to give quite comparable results (Borja et al. 2011, Borja et al. 2007). The use of the latter option has several practical advantages: (a) it can be calculated, automated and modified more straightforwardly than a multivariate method and (b) it gives understandable sub-EQR scores for the indicators used which is very informative for the specialist user and (c) these sub-EQR scores can be used well for communication with water managers and policy makers. Therefore, it is concluded that the use of a linear combination of the Shannon index and the selected AMBI is the most useful EQR normalization method for the optimized BEQI.

It appears that two-point calibration using a reference sample and corresponding indicator values (EQR = 1) and a theoretical bad sample (EQR= 0) is the commonly used method. The Swedish method to look for a good/moderate boundary in a pressure/impact curve is ecologically very relevant (Josefson et al. 2009) but is less used; probably because it is difficult to find a complete pressure gradient and benthic impact dataset. The finding of real reference samples is a fundamental problem; because these samples would probably have to originate from before the advent of the large steam trawlers in the North Sea around 1880 (De Vooys et al. 2004). To solve this problem, the best sample, or pool of samples, from a large dataset from the same ecotope (at least salinity zone) is selected as the reference sample. The UK-regression method to estimate reference values corrected for salinity appears to be a useful statistical method. This method is based on indicator values of individual samples. Considering the large and long-term effects which fisheries have had on the marine benthic communities in the past and present, this is still considered to give an underestimation of the quality of marine benthic communities.

An important advantage of both the Shannon index and the AMBI is that reference values consist of a single value, instead of an extensive dataset as in the current BEQI. Furthermore, there are many reference values already available from the intercalibration process, which will be used as a starting point for the calibration of the new Dutch BEQI in the coming (inter-) calibration project. Recently, a reference estimation model has been made available by the UK in the intercalibration process (Phillips, pers. comm.).

12. Design of optimized BEQI. It appears logical and optimal to assess benthic communities with a combination of the Shannon index and an AMBI-like categorisation of sensitive/opportunistic species including categories worked out for typical Dutch pressures (bottom fisheries, extraction, etc), in a formula comparable to the DKI and NQI. It is recommended to calculate an average score of the three AMBIs, which emphasizes the lowest (most sensitively reacting) AMBI score per water body/ecotope. The use of equal weight factors for H', and the most sensitive AMBI, is recommended (a) in order to give in principle equal weight to the diversity component and the sensitive/opportunistic species component of the metric and (b) in view of the good intercalibration results obtained with the BKI and NQI so far. It is possible to add a correction factor for the total number of individuals using the subformula from the NQI/BQI; it will be checked if this is necessary in the (inter)calibration project. This leads to the following optimized BEQI level 3 metric:

$$\text{BEQI level 3} = 0.5 *((1 - \text{AMBI}_{ass}/7) /(1 - \text{AMBI}_{ref}/7)) + (0.5 *(H'_{ass}/H'_{ref}))$$

Correction factor option 1: $* (N_{tot}/N_{tot} + 5)$
Correction factor option 2: specify a minimum amount of N_{tot}
Please note that when $N_{tot} = 100$, the correction factor in option 1 is < 5% which becomes negligible) The choice between these two options will be made in the coming (inter-) calibration project.

Note: if necessary the weight factors for H' and AMBI in the proposed new metric can be adjusted in order to obtain metric results which are in agreement with Dutch expert judgement.

It is also recommended to routinely co-calculate the indicators Species Richness (S) Evenness (E) as supporting information, not for WFD assessment, but to interpret changes in H'. This is however not necessary if S is added to the BEQI metric.

13. AMBI validation. The AMBI species sensitivity classes are currently validated and expanded to fisheries and extraction/sedimentation in a currently finalized Gimaris/Waterdienst project.

14. Calibration and intercalibration. An (inter-)calibration project is planned in January-March 2011 to calibrate the new BEQI with RWS monitoring data, and to calculate EQR-data for the Westerschelde and other relevant EU water bodies from the common database.

5.3 Other recommendations

Sampling and analysis

15. Sampling area. It is necessary to sample at least 0.19 m^2 per ecotope to be assessed. This guideline for sampling has been agreed on in the NEA GIG intercalibration working group. This minimum sampling area is necessary to obtain sufficiently reliable assessment results.

16. Species identification. In the past RWS MWTL marine benthos samples have not been analysed sufficiently down to the species level in two of the three major marine regions in the Netherlands (Gittenberger 2011). Therefore, it is very important that the Dutch marine benthos analysis is improved, in order to obtain identifications down to the species level for practically all common species. This improvement of the analytical method is currently undertaken by the RWS Waterdienst.

17. Since epibenthic scavenger and predator species (e.g. crustaceans, gastropods and sea stars) are important indicator species for bottom trawl fisheries (Rumohr 2000), it is recommended to co-analyze the epibenthos into the WFD assessment of macrozoobenthos in water bodies where bottom fisheries are a major pressure (which is pretty much everywhere on the continental shelf).

Assessment

18. Data pooling. It is in principle necessary to use a randomization procedure to select the most representative sample pool, if more than the standardized amount of 0.19 m^2 of sample area is available. In the intercalibration process however, a pragmatic choice has been made to pool the first samples in the database which add up to 0.19 m^2.

For the estimation of reference values, the samples from a specific water body/ecotope with the highest Shannon and AMBI scores, respectively, will be pooled to 0.19 m^2 in order to obtain reference values for the indicator H' and AMBI for the specific water body/ecotope.

19. Calibration/intercalibration process. It is recommended to analyse (a) the intercalibration common database and (b) other useful Dutch marine benthos data in order to (1) check the sensitivity of the new BEQI and (2) obtain suitable reference data using an essential (not unnecessary detailed) ecotope classification and (c) perform initial tests with the pressure index method (Borja et al. 2011). This data project will be performed in spring 2011.

20. Human pressure calibration. It is recommended to use the semi-quantitative pressure index method of Borja et al. (2011) for the Dutch marine waters. It can be checked if other methods to calculate the combined pressure index may lead to further improved results.

21. Application of the new BEQI. It is recommended to study the application of the new BEQI in large marine projects such as the Voordelta monitoring project for compensation of the second Maasvlakte, and in the Marine Strategy Framework Directive.

6 References

Aakerman J, Klamer C, Schipper , Bakker J, Bellert B, Pijnenburg J. (2004). Stoffen in de Noordzee en de Nederlandse kustzone in 2003 - ftalaten, vlamvertragers, organotin- en geperfluoreerde verbindingen en effectgerichte metingen. Report RIKZ/2004.040, RIKZ, Den Haag, The Netherlands.

Armonies W. (2000). On the spatial scale needed for benthos community monitoring in the coastal North Sea. *Journal of Sea Research*, 43: 121–133.

AZTI, 2010, Instructions for the use of the AMBI index software (version 4.0), http://ambi.azti.es/

Bergman MJN, Van Santbrink JW (2000). Mortality in megafaunal benthic populations caused by trawl fisheries on the Dutch continental shelf in the North Sea in 1994. *ICES Journal of Marine Science*, 57:1321–1331

Beukema JJ (1988). An evaluation of the ABC-method (abundance/biomass comparison) as applied to macrozoobenthic communities living on tidal flats in the Dutch Wadden Sea. *Marine Biology*, 99:425–433, .

Birklund J, Wijsman JWM (2005). Aggregate extraction: a review of the effect on ecological functions. Technical Report Z3297.10, DHI Water and Environment; WL/Delft Hydraulics.

Borja A, Bald J, Franco J , Larreta J , Muxika I, Revilla M, Rodrigues JG, Solaun O, Uriarte A, Valencia V (2009). Using multiple ecosystem components in assessing ecological status in Spanish (Basque Country) Atlantic marine waters. *Marine Pollution Bulletin* 59: 54–64.

Borja A., Dauer DM (2008). Assessing the environmental quality status in estuarine and coastal systems: Comparing methodologies and indices. *Ecological Indicators* 8: 331–337.

Borja A, Josefson AB, Miles A, Muxika I, Olsgard F, Phillips G, Rodriguez JG, Rygg B (2007). An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. *Marine Pollution Bulletin* 55: 42–52.

Borja A, Muxika I, Franco J (2000). The application of a Marine Biotic Index to different impact sources affecting soft-bottom benthic communities along European coasts. *Marine Pollution Bulletin* 40-12: 1100–1114.

Borja A, Barbone E, Basset A, et al.(2011). Response of single benthic metrics and multimetric methods to anthropogenic pressure gradients, in five distinct European coastal and transitional ecosystems (accepted for publication). Doi: 10.1016/j.marpolbul.2010.12.009.

Brinkman AG, Jansen JM (2007). Draagkracht en exoten in de Waddenzee. Technical Report C073/07, Imares.

Connell H (1978). Diversity in tropical rain forests and coral reefs. *Science* 1999: 1302–1310.

Craeymeersch JA, Hamerlynck O, Hostens K, Vanreusel A,Vincx M (1990). De ekologische ontwikkeling van de Voordelta, deelrapport 1. de huidige ekologische situatie van de Voordelta. Technical report, Delta Instituut voor Hydrobiologisch Onderzoek.

Craeymeersch JA, Escavarage V (2010). Jaarrapport 2010 PMR Monitoring Natuurcompensatie Voordelta, deel B perceel benthos (2010).

Dauvin JC, Ruellet T (2007) Polychaete/amphipod ratio revisited. *Marine Pollution Bulletin* 55: 215–224.

Dauvin JC, Bellan G, Bellan-Santini D (2010). Benthic indicators: From subjectivity to objectivity - Where is the line? *Marine Pollution Bulletin* 60: 947–953.

Dekker R (1989). The macrozoobenthos of the subtidal western Dutch Wadden Sea. I. Biomass and species richness. *Netherlands Journal of Sea Research* 23: 57–68.

Devlin MJ, Barry J, Mills DK, Gowen RJ, Foden J, Sivyer D, Tett P (2008). Relationships between suspended particulate material, light attenuation and Secchi depth in UK marine waters. *Estuarine, Coastal and Shelf Sciences* 79: 429–439.

Duineveld GCA, Bergman MJN, Lavaleye MSS (2007). Effects of an area closed to fisheries on the composition of the benthic fauna in the southern North Sea. *ICES Journal of Marine Science* 64: 1–10.

EU (2007). WFD intercalibration technical report, Part 3: Coastal and Transitional Waters, Section 2: Benthic Invertebrates.

Faber W et al. (2011). Richtlijn KRW Monitoring Oppervlaktewater en Protocol Toetsen & Beoordelen (Dutch standard methods for WFD monitoring and assessment).

Fleischer D, Grémare A, Labrune C, Rumohr H, Vanden Berghe E, Zetter ME (2007). Performance comparison of two biotic indices measuring the ecological status of water bodies in the Southern Baltic and Gulf of Lions. *Marine Pollution Bulletin* 54: 1598-1606.

Gittenberger A, van Loon WMGM (2011). A list of common, marine, benthic, animal species of the Netherlands and their properties. Gimaris report 2011.01.

Grémare A, Labrune C, Vanden Berghe E, et al. (2009). Comparison of the performances of two biotic indices based on the MacroBen database. *Marine Ecology Progress Series* 382: 297–311.

Heink U, Kowarik I (2010). What are indicators? On the definition of indicators in ecology and environmental planning. *Ecological indicators* 10: 584–593.

Herman PMJ, Middelburg JJ, van de Koppel J, Heip CHR (1999). Ecology of estuarine macrobenthos. *Advances in Ecological Research* 29: 195-240

Hurlbert SH (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52: 577-586.

Jennings S, Dinmore TA, Duplisea DE, Warr KL, Lancaster JE (2001). Trawling disturbance can modify benthic production processes. *Journal of Animal Ecology* 70: 459–475.

Jongman RHG, Jager JC, Looman CWN, Ter Braak CJF, Van Tongeren OFR, Burrough PA, Van de Nes TJ (1995). Data analysis in community and landscape ecology. Cambridge University Press.

Josefson B, Blomqvist AM, Hansen JLS, Rosenberg R, Rygg B (2009). Assessment of marine benthic quality change in gradients of disturbance: Comparison of different Scandinavian multi-metric indices. *Marine Pollution Bulletin* 58: 1263–1277.

Kabuta S (2000). Graadmeters voor de Noordzee, RIKZ rapport.

Knoben R, Kamsma R (red) (2004) Achtergronddocument referenties en maatlatten voor macrofauna. STOWA Report, concept version Dec. 2004.

Lavaleye MSS (1999). GONZ rapport, Rapport Graadmeters van de Noordzee - Infaunal Trophic Index (ITI) van het macrobenthos en structuur macrobenthos gemeenschap (verhouding r- en K-strategen). Technical report, NIOZ, Den Burg, The Netherlands.

Lindeboom HJ, De Groot SJ (1998). Impact II - the effects of different types of fisheries on the North Sea and Irish Sea benthic ecosystems. Technical report, NIOZ, Den Burg, IJmuiden, The Netherlands.

Loewe P (ed.) (2009). System Nordsee - Zustand 2005 im Kontext langzeitlicher Entwicklungen. Technical Report Nr. 44/2009, Bundesamtes für Seeschifffahrt und Hydrographie, Hamburg, Germany.

Meire PM, Dereu J (1990). Use of abundance/biomass comparison method for detecting environmental stress: some considerations based on intertidal macrozoobenthos and bird communities. *Journal of Applied Ecology* 27: 210–223,

Meyer Th, Berg T, Fürhaupter K, Maczassek K (2007). Bewertung von Makrozoobenthos im Küstengewässer der FGE Weser und Elbe nach WRRL - Test einer MarBIT-Erweiterung. Technical report, MARILIM Gewässeruntersuchung, Kiel, Germany.

Moulin L, Caterino AI, Claessens T, Dubois P (2011). Effects of seawater acidification on early development of the intertidal sea urchin *Paracentrotus lividus* (Lamarck 1816). *Marine Pollution Bulletin* 62: 48-54.

Mugurran AE (1988). Ecological diversity and its measurement. Chapman and Hall, London, UK.

Muxika I, Borja A, Bonne W (2005). The suitability of the marine biotic index (AMBI) to new impact sources along European coasts. *Ecological Indicators* 5: 19–31.

Muxika I, Borja A, Bald J (2007). Using historical data, expert judgement and multivariate analysis in assessing reference conditions and benthic ecological status, according to the European Water Framework Directive. *Marine Pollution Bulletin* 55: 16–29.

Muxika I, Somerfield PJ, Borja A, Warwick RM (2010).Assessing proposed modifications to the AZTI marine biotic index (AMBI). ICES CM 2010/H:07.

Pearson TH, Rosenberg R (1978). Macrobenthic succession in relation to organic enrichment and pollution of the marine environment. *Oceanography and Marine Biology: An Annual Review* 16: 229–311.

Pinto R, Patrício J, Baeta A, Fath BD, Neto JM (2009). Review and evaluation of estuarine biotic indices to assess benthic condition. *Ecological Indicators* 9: 1-25.

Rees H, Eggleton JD, Rachor E, Vanden Berghe E (Eds.) (2007). Structure and dynamics of the North Sea benthos. Technical Report No. 288, ICES.

Reiss H, Kröncke I (2005). Seasonal variability of benthic indices: An approach to test the applicability of different indices for ecosystem quality assessment. *Marine Pollution Bulletin* 50: 1490–1499.

Rumohr H, Kujawski T. (2000). The impact of trawl fishery on the epifauna of the southern north sea. *ICES Journal of Marine Science*, 57: 1389–1394.

Rosenberg R, Blomqvist M, Nilsson HC , Cederwall H, Dimming A (2004). Marine quality assessment by use of benthic species-abundance distributions: a proposed new protocol within the European Union Water Framework Directive. *Marine Pollution Bulletin* 49: 728–739.

Roxburgh SH, Shea K, Wilson JB (2004). The Intermediate Disturbance Hypothesis: Patch Dynamics and mechanisms of species coexistence. *Ecology* 85: 359–371.

Salas F, Neto JM, Borja A, Marques JC (2004). Evaluation of the applicability of a marine biotic index to characterize the status of estuarine ecosystems: the case of Mondego estuary (Portugal). *Ecological Indicators* 4: 215–225.

Sanders HL (1968). Marine benthic diversity: a comparative study. *American Naturalist* 102: 243-282.

Schroeder A (2003). *Community dynamics and development of soft bottom macrozoobenthos in the German Bight (North Sea) 1969 - 2000.* PhD thesis, University of Bremen, Germany.

Smeets E, Weterings R (1999). Environmental indicators: Typology and overview. Technical Report No. 25, EEA.

Steenbergen J (2004). Het effect van sterk wisselende zoutgehalten op het benthos in de Westerschelde en de Haringvlietmonding. Technical Report C075/04, RIKZ.

Tsimplis MN, Shaw AGP, Flather RA, Woolf DK (2006). The influence of the North Atlantic oscillation on the sea-level around the northern European coasts reconsidered: the thermosteric effects. *Philosophical Transactions of the Royal Society A*, 364: 845–856, .

Twisk F, Van Loon W, Ysebaert T, Herman P (2009). KRW-beoordeling van macrofauna in kust- en overgangswateren met BEQI. *H2O* 19: 35–37.

Van der Molen DT, Pot R (Eds) (2006). Referenties en maatlatten voor overgangs- en kustwateren ten behoeve van de Kaderrichtlijn Water, Update 2006. STOWA Report no. 2004-44.

Van Hoey G, Drent J, Ysebaert T, Herman P (2007). The Benthic Ecosystem Quality Index (BEQI), intercalibration and assessment of Dutch coastal and transitional waters for the Water Framework Directive. Technical Report NIOO 2007-02, NIOO, Yerseke; NIOZ, Texel, The Netherlands.

Wilson JB (1994). The 'Intermediate Disturbance Hypothesis' of species coexistence is based on patch dynamics. *New Zealand Journal of Ecology* 18: 176–182.

Ysebaert T, de Mesel I , Herman P (2008). Kaderrichtlijn Water - Achtergronddocument Zoute Macrofauna 2008. Technical Report C076/08, Wageningen Imares, The Netherlands.

De Vooys CGN, R Dapper, J van der Meer, MSS Lavaleye, H.J. Lindeboom (2004). Het macrobenthos op het Nederlands continentale plat in de Noordzee in de periode 1870-1914 en een poging tot vergelijking met de situatie in de periode 1970-2000, NIOZ report 2004-2.

Vos P (2006). Statistische evaluatie van het macrofauna meetnet in de Rijkswateren, CML Rapport (in Dutch), Leiden, The Netherlands.

Ware SJ, Rees HL, Boyd SE, Birchenhough SN (2009). Performance of selected indicators in evaluating the consequences of dredged material relocation and marine aggregate extraction. *Ecological Indicators* 9: 704–718.

Warwick RM, Clarke KR (1996). Relationships between body-size, species abundance and diversity in marine benthic assemblages: facts or artefacts? *Journal of Experimental Biology and Ecology* 202: 63–71.

WFD-UKTAG (2008). UKTAG coastal water assessment method: benthic invertebrate fauna. Invertebrates in soft sediments (infaunal quality index (IQI)). Sniffer, Edinburgh, Scotland, UK.

Word JQ (1978). The infaunal trophic index, in: Bascom, W. (Ed.), Southern California Coastal Water Research Project Annual Report, El Segundo, California, USA, pp. 19-39.