Variability in the assignment of maturity stages of plaice (*Pleuronectes platessa* L.) and whiting (*Merlangius merlangus* L.) using macroscopic maturity criteria

H.D. Gerritsen^{a,b,*} and D. McGrath^b

^a Fisheries Science Service, Marine Institute, Galway Technology Park, Parkmore, Galway, Ireland

^b Commercial Fisheries Research Group, Department of Physical and Life Sciences, Galway Mayo Institute of Technology, Dublin Road, Galway, Ireland

* Corresponding author. Tel: +353 91 780 368; Fax: +353 91 730 470; Email address: hans.gerritsen@marine.ie

(Correspondence address: H.D. Gerritsen, Fisheries Science Service, Marine Institute, Galway Technology Park, Parkmore, Galway, Ireland)

Abstract

This study investigates if a macroscopic maturity scale can be applied consistently, by examining the variability between and within ten people who repeatedly assessed the sex and maturity stages of 80 plaice (*Pleuronectes platessa* L.) and 79 whiting (*Merlangius merlangus* L.) gonads. In most cases, agreement within assessors was not significantly higher than agreement between assessors, suggesting that variability was random and not due to differences in interpretation. This finding was supported by the fact that a significant bias was only found for one assessor. Some maturity stages were assigned quite consistently, while other stages were not defined objectively enough to be assigned reliably, even when fish were assessed repeatedly by the same person. For both species, well-defined maturity scales with fewer stages would be preferable over scales that distinguish a larger number of maturity stages. As maturity staging will always contain a form of subjective judgement, it should be subject to continuous quality control measures.

Keywords: maturity scale; reproduction; operator error; operator bias; quality control; quality assurance.

1. Introduction

Operator error is potentially a significant factor in many aspects of data collection in fisheries science. While any measurement error should be subject to some sort of quality control mechanism whereby accuracy, precision and bias are being monitored; this is particularly important when subjective criteria are used to quantify a variable. Although even measures such as fish length can have a source of operator error (King, 1995), it is more likely to be a problem for parameters like fish age and maturity stage, where subjective judgements play a large role. There is a large body of literature available on the quality control and quality assurance in the field of age

reading (e.g. Kimura and Lyons, 1991; Campana et al., 1995; Eltink et al. 2000; Campana 2001; Kimura and Anderl, 2005). However, the literature on the assignment of maturity stages to fish gonads is more limited. Most work seems to be concerned with routine validation of maturity staging. The most widely used method to determine the maturity stage of fish is visual examination of the gonads, applying macroscopic criteria, a selection of which is usually validated by histological analysis (West, 1990). This work is published mostly as working documents or reports, with a few exceptions (e.g. Tomkiewicz et al. 2003; Claereboudt et al. 2005). However, apart from validation (a measure of accuracy), any quality control program should also include measures of precision and bias. Measures of precision should include variability between as well as within assessors, which can only be obtained by repeated assessment of the same sample. Bias measures whether disagreement is systematic and differences between workers are the result of differences in interpretation of the maturity scale or whether disagreement is random. This information can be used to improve the design of maturity scales and to monitor training or calibration exercises.

It appears that only one study has been published in which precision of maturity staging was investigated (ICES, 2005). In this study, 72 images of herring gonads were sent to a number of laboratories to be assessed by various people. The study concluded that agreement was higher for maturity scales that distinguished fewer stages, but no correction was made for the fact that a scale with fewer stages has a higher amount of chance agreement. No comparisons were made within assessors nor was it investigated whether disagreement was random or systematic. Additionally, the use of photographic material might not accurately reflect the real situation.

The current study investigates whether macroscopic maturity scales can be applied consistently by examining the variability in the assignment of maturity stages to fresh samples of plaice (*Pleuronectes platessa* L.) and whiting (*Merlangius merlangus* L.). The study will address the agreement of individual maturity stages, agreement between and within workers and check whether disagreement is random or systematic. Histological validation will not be addressed in the current study but will be the subject of future work.

2. Methods

Two samples of plaice were taken during the Irish Groundfish Survey, carried out in 2004 by the Marine Institute. Additionally, two samples of whiting were taken during the Biological Sampling Survey, which was undertaken in 2005 by the Marine Institute. The first plaice sample was taken on 17 October in ICES area VIa-South (north of Ireland); the second sample was taken on 4 November in ICES area VIIg (Celtic Sea). Both samples were selected from a catch of about 150 plaice, which were sorted into the available maturity categories. Around seven fish were chosen at random from each available sex and maturity category, so that the sample contained approximately equal numbers in each category. Both plaice samples contained a total of 40 fish. At the time of sampling, many plaice had already started gonad development for the next spawning season, which starts around January (Armstrong et al., 2001), so a variety of maturity stages were available. Both samples of whiting were taken in ICES area VIIj (Celtic Sea) on 4 March, just before the peak of

spawning for whiting. During that time nearly all maturity stages except spent fish were present. The whiting samples were taken from a catch of around 300 fish and contained 40 and 39 fish respectively from the range of available maturity stages. The fish were tagged with complex labels to ensure the assessors would not remember individual fish. The assessors were not aware that there were approximately equal numbers of fish from each maturity stage.

The maturity of the fish was assessed using the maturity scale given in Table 1. In both plaice samples, maturity stages F1, F2 and F4 were identified for females and stages M1, M3 and M4 were found for males. For plaice, maturity stages F3 and M2 are usually not distinguished but the numbering is maintained for consistency with other species. In the whiting samples, maturity stages F1 to F5 were identified for females and stages M1 to M4 for males.

The plaice samples were assessed by four people each to allow for comparisons between workers. The samples were then stored overnight at 4°C in a sealed plastic bag to keep them as fresh as possible. The next day, the samples were assessed again by the same set of people to allow for comparisons within, as well as between, assessors. The whiting samples were assessed by three people, both samples were assessed twice by each person on the same day; these samples could not be stored overnight as they tended to deteriorate more rapidly than plaice. There was some variation in the experience of the workers concerned. Person A had a number of years of experience in the assessment of maturity stages trained all others except person C, who was trained independently. All others had little previous experience in maturity staging. Training took place while at sea using both fresh and photographic material.

To examine the agreement between two people classifying the same sample, Cohen (1960) proposed a coefficient of agreement, kappa, which takes into account that a certain amount of agreement between two judges might be expected by chance. Cohen's kappa is, in other words, the proportion of agreement after chance agreement has been removed. It is given by:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (1)$$

where p_o is the observed proportion of agreement and p_e is the proportion of agreement expected by chance. For a comparison between persons A and B using n categories ($i = 1, 2, \dots, n$), the proportion of agreement expected by chance is given by:

$$p_e = \sum_{i=1}^n p_{a,i} \cdot p_{b,i} \quad (2)$$

where $p_{a,i}$ is the proportion of the sample that person A attributes to category i and $p_{b,i}$ is the proportion of the sample that person B attributes to category i . Kappa ranges from -1 to 1, where $\kappa = -1$ indicates complete disagreement and $\kappa = 1$ indicates complete agreement. Cohen's kappa is widely used in the social and medical sciences where nominal and ordinal scales are used regularly.

The interpretation of kappa can be sensitive to asymmetric or systematic disagreement and to very high or very low prevalence (Schuster, 2004). Systematic disagreement will be tested for using binomial tests (see below). No cases of extreme prevalence occurred as the samples were selected to have approximately equal numbers in each maturity category. Additionally, the use of kappa requires the assumption of independent ratings (Cohen, 1960). This assumption is met, as each sample only contributes to one paired rating for each comparison and the assessors worked without knowledge of the others' results. The fact that two assessors assess the same samples does not contravene the assumption of independence. Cohen's kappa was originally proposed for the use of nominal scales and another parameter, weighted kappa, was later introduced to deal with ordinal scales (Cohen, 1968). The maturity scale is an ordinal scale as one maturity stage follows the next, however for the current study no distinction will be made between a disagreement over consecutive points on the scale (e.g. F1-F2) or points further apart (F1-F5), therefore unweighted kappa is appropriate in the present case.

Differences between correlated kappa values can be evaluated by bootstrapping the pairs of observations (Efron and Tibshirani, 1993; McKenzie et al., 1996). For each kappa value, 1000 bootstrap replications were performed and confidence intervals (CI) were generated for the differences between two bootstrapped kappa values. The 95% CI were estimated from the percentiles of the differences. If the CI did not include zero, the difference was considered larger than zero and kappa values were significantly different at the 5%. For differences between two mean kappa values, the means were calculated for each bootstrap replication and otherwise the same procedure was followed.

Kappa quantifies agreement but it does not give any information on whether disagreement is random due to bias. One assessor might interpret the maturity scales in a systematically different way from the others or disagreement might be random. If disagreement is random, the number of fish that one assessor assigned to a certain category with which a second assessor did not agree, should not be significantly different from the number of fish that the second assessor assigned to this category with which the first did not agree. If this is not the case, disagreement could be systematic and might result from a difference in the interpretation of the maturity scale. As the number cases of disagreement in each category is generally low ($n < 10$), a binomial test is appropriate to test for a systematic difference (Zar, 1999). The binomial test was performed separately for each of the available categories. To reduce the risk of type I error (false positive), a Bonferroni correction for repeated tests was applied to the significance level (Bland and Altman, 1995). With k categories (one test for each category) and an overall significance level of 5%, the significance level for each individual test, α_k , is given by:

$$\alpha_k = 0.05/(k - 1) \quad (3)$$

All computations were performed in the R environment (Venables et al., 2005)

3. Results

Table 2 shows the total of all pair-wise combinations between assessors by maturity stage for the plaice samples. There was some variability in the assignment of male and female plaice at the early maturity stages. This was mainly due to one person, assessor G. Stage F4 was assigned with the highest consistency in plaice, while there was considerable disagreement between stages F1 and F2. For male plaice most of the disagreement was between stages M3 and M4. Table 3 shows the figures for whiting. The most consistent maturity stage for whiting was F5, with high disagreement between F2 and F3 for females and around M2 and M3 for males. In general, most of the disagreement was confined to consecutive maturity stages, indicating that the distinction between these stages was not clearly defined.

Kappa values were calculated for comparisons between and within assessors (Tables 4 and 5). To test if there were differences in agreement between the first and second time the samples were assessed (e.g. due to deterioration of the samples), the mean kappa values for all comparisons within the first assessment, were compared with those within the second. For plaice sample I, the mean kappa value was 0.83 for the first time and 0.76 for the second time the samples were assessed. For plaice sample II, mean kappa values were 0.70 for the first and 0.73 for the second assessment. Neither of these differences was significant (bootstrapped difference, $p > 0.05$). Whiting sample I had a higher mean kappa value for the first assessment (0.71) than for the second (0.60) but whiting sample II had a lower mean kappa value for the first assessment (0.65) than for the second (0.80). Again, neither difference was significant at the 5% level. Therefore, there seems to be no reason to assume that there were any differences in agreement between the first and the second time the samples were assessed.

To examine if agreement within assessors was higher than between assessors, the differences in the mean kappa values for comparisons within and between assessors were evaluated. For plaice sample I, the mean kappa value for comparisons within assessors was significantly higher than the overall mean kappa value between assessors (0.89 and 0.80 respectively; bootstrapped difference: $p < 0.05$). For plaice sample II, the mean kappa within assessors nearly the same as the mean kappa between assessors (0.70 and 0.71); the difference was not significant. As both whiting samples were assessed by the same set of people, the samples were combined; the mean kappa within assessors was higher than between assessors (0.70 and 0.65 respectively) but this difference was not significant.

The most experienced workers, A and C, obtained higher kappa values for within-assessor comparisons than most others, although assessors D and F also scored highly (Table 4). Person C scored low kappa values for between-assessor comparisons and appeared to assign some maturity stages in a different way from the others: for the first time plaice sample I was assessed, a significant systematic difference was found between persons A and C (two-tailed binomial test, $p < 0.01$), for the second time the sample was assessed systematic differences found between B and C and between C and D (two-tailed binomial test, $p < 0.01$). In every case, person C was more likely to assign stage F2 and less likely to assign F1 than the others. Assessor C can therefore be considered to be biased towards the others. No significant systematic differences were found in any of the other comparisons. Another person that deviated from the general pattern was assessor G (plaice sample II). Agreement was very low for all

comparisons with this assessor and particularly within person G. This indicates that person G was generally inconsistent in the assignment of maturity stages.

4. Discussion

An important assumption of the study is that the first time the samples were assessed, was independent from the second assessment. Although the workers were aware that they were assessing the same sample twice, the experiment was designed to make it difficult for anyone to remember individual fish. The fish were labelled with complex codes and samples contained few very small, very large or otherwise distinctive fish. The two whiting samples were given in a different order than the assessors were told to expect but this was not noticed by anyone so individual codes or fish were not remembered. The first and second assessment of each sample can therefore be considered independent. It is also implicitly assumed that the condition of the samples did not deteriorate over time. There is little reason to assume that they did, as the mean agreement did not decrease significantly for any of the plaice or whiting samples. It can therefore be assumed that comparisons between and within the assessors are valid.

Variability appeared to be high for certain maturity stages, while others appeared to be quite well defined. It is assumed that fish of either sex at maturity stages 1 and 2 will not spawn in the current season and that all fish of stage 3 and higher are mature. Therefore the most important distinction is between stages 2 and 3. For plaice there was little disagreement between the assignment of mature and immature fish. However, for whiting, there was considerable variability between the immature females (F2) and maturing females (F3). This could potentially have a large effect on the estimate of the proportion of mature fish in the population. In the present case, maturity stage F3 was quite rare in the catches (<4%), so this effect is probably small for the current survey. However, these results highlight the importance of timing of sampling. Size frequency analysis of the oocytes of plaice and whiting (Gerritsen et al., unpublished data) shows that vitellogenic cells in plaice quickly reach a relatively large size. On the other hand, in whiting there is no obvious gap in the size range between pre-vitellogenic and vitellogenic cells. It is therefore more difficult to distinguish females with vitellogenic cells (F3; mature) from those without vitellogenic cells (F2; immature) in whiting.

One person, assessor C, was shown to be biased towards the other assessors. It is interesting to note that C was trained independently from the others. If person C is omitted from plaice sample I, mean kappa within and between assessors is no longer significantly different. So, with the exception of person C, there does not seem to be a higher agreement within assessors than between them. This suggests that the disagreement found was almost entirely due to chance and not due to differences in interpretation of the maturity scale. This has a number of implications, firstly, it is unlikely that the consistency of most assessors can be much improved by further training; some maturity stages simply could not be assigned consistently, even when samples were assessed repeatedly by the same person. On the other hand, it should be possible to address the issues of systematic disagreement (assessor C) and major inconsistency (assessor G) by calibration exercises and additional training.

For the purposes of quality control, it would be very useful to repeatedly assess a number of maturity samples each time they are collected. However comparing more than two assessors or repeatedly assessing the same assessor is quite time consuming and might not be possible on a routine basis. Additionally, unlike age reading, it will not be possible to return to the maturity samples and assess them again if agreement falls below a certain threshold. On the other hand it would be feasible for two assessors to note maturity stages independently of each other as part of the normal sampling procedure. This information would be valuable to monitor the progress of training or calibration exercises. It would also contribute to finding a realistic value of kappa at which agreement is considered adequate. At the moment this is not possible, as too few samples have been assessed so far.

Hunter and Macewicz (2003) suggested that maturity scales could be improved by reducing the number of classes and focussing on the most reliable characteristics. The present study seems to confirm this; maturity stages that were not unequivocally defined could not be distinguished consistently and therefore the value of a maturity scale with a large number of stages is doubtful. The study on herring gonads (ICES, 2005) seems to support this finding. A maturity scale that distinguishes between virgin, maturing, spawning and spent fish might be a good compromise between gleaned detailed information and obtaining consistent results. The current findings will most likely generalise to other species in other areas as most maturity scales are based on similar criteria.

Acknowledgements

We thank all the people who contributed their precious time during the Irish Groundfish Survey 2004 and the Biological Sampling survey 2005: Robert Bunn, Afra Egan, Imelda Hehir, Graham Johnston, Kieran McCann, Ayesha Power, Gavin Power and Turloch Smith. We also thank Dr. Tasman Crowe, Zoology Department, University College, Dublin for valuable discussions on the experimental design and two anonymous reviewers for their helpful comments.

References

Armstrong, M.J., Connolly, P., Nash, R.D.M., Pawson, M.G., Alesworth, E., Coulahan, P.J., Dickey-Collas, M., Milligan, S.P., O'Neil, M.F., Witthames, P.R., Woolner, L., 2001. An application of the annual egg production method to estimate the spawning biomass of cod (*Gadus morhua* L.), plaice (*Pleuronectes platessa* L.) and sole (*Solea solea* L.) in the Irish Sea. ICES J. Mar. Sci. 58, 183-203.

Bland, J.M., Altman, D.G., 1995. Multiple significance tests: the Bonferroni method. Br. Med. J. 310, 170.

Bromley, P.J., 2000 Growth, sexual maturation and spawning in Central North Sea plaice (*Pleuronectes platessa* L.), and the generation of maturity ogives from commercial catch data. J. Sea Res. 44, 27-43.

Campana, S.E. 2001. Accuracy, precision and quality control in age determination, including a review of the use and abuse of age validation methods. *J. Fish. Biol.*, 59, 197-242.

Campana, S.E., Annand, M.C., McMillan, J.I., 1995. Graphical and statistical methods for determining the consistency of age determinations. *Trans. Am. Fish. Soc.* 124, 131-138.

Claereboudt, M.R., McIlwain, J.L., Al-Oufi, H.S., Ambu-Ali, A.A. 2005 Patterns of reproduction and spawning of the kingfish (*Scorberomorus commerson*, Lacépède) in the coastal waters of the Sultanate of Oman. *Fish. Res.*, 73, 273-282.

Cohen, J., 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37-46.

Cohen, J., 1968. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 70, 213-220.

Efron, B., Tibshirani, R.J., 1993. An introduction to the bootstrap. Chapman & Hall, Boca Raton.

Eltink, A.T.G.W., Newton, A.W., Morgado, C., Santamaria, M.T.G., Modin, J., 2000. Guidelines and tools for age reading comparisons. European Fish Ageing Network (EFAN), Flødevigen.

Hunter, J.R., Macewicz, B.J., 2003. Improving the accuracy and precision of reproductive information used in fisheries. In: Kjesbu, O.S., Hunter, J.R., Witthames, P., (Eds.), Report of the working group on modern approaches to assess maturity and fecundity of warm- and cold-water fish and squids. *Fisken og havet*, 12-2003, Institute for Marine Research, Bergen, pp. 57-68

ICES, 2005. Report for the Planning Group for Herring Surveys 2005. ICES CM 2005/G:04. International Council for Exploration of the Sea (ICES), Bergen.

Kimura, D.K., Anderl, D.M., 2005. Quality control of age data at the Alaska Fisheries Science Center. *Mar. Freshwater Res.* 56, 783-789

Kimura, D.K., Lyons, J.J., 1991. Between-reader bias and variability in the age-determination process. *Fish. Bull.* 89, 53-60.

King, M., 1995 Fisheries biology, assessment and management. Blackwell Science, Oxford.

McKenzie, D.P., Mackinnon, A.J., Péladeau, N., Onghena, P., Bruce, P.C., Clarke, D.M., Harrigan, S., McGorry, P.D., 1996. Comparing correlated kappas by resampling: is one level of agreement significantly different from another? *J. Psychiat. Res.* 30, 483-492.

Schuster, C., 2004. A note on the interpretation of weighted kappa and its relations to other rater agreement statistics for metric scales. *Educ. Psychol. Meas.*, 62, 243-253.

Tomkiewicz, J., Tybberg, L., Jespersen, Å., 2003. Micro- and macroscopic characteristics to stage gonadal maturation of female Baltic cod. *J. Fish Biol.* 62, 253-275.

Venables, W.N., Smith, D.M., R-Development-Core-Team, 2005. An Introduction to R. (Notes on R: A programming environment for data analysis and graphics.) Version 2.1.1.

West, G., 1990. Methods of assessing ovarian development in fishes: a review. *Aust. J. Mar. Freshwater. Res.* 41, 199-222.

Zar, J.H., 1999. Biostatistical analysis, Prentice-Hall, New Jersey.

Tables

Table 1. Macroscopic maturity scale for plaice and gadoids.

Females		Males	
F1	Ovary small, contents translucent, nearly colourless	M1	Testes tiny, translucent and undeveloped
F2	Contents blurred translucent, pinkish, no oocytes visible	M2	Gadoids: lobules developing but still pinkish and translucent. Plaice: no stage M2
F3	Gadoids: ovary opaque, contents yellow, individual oocytes not yet visible. Plaice: no stage F3	M3	Testes developing, creamy white and opaque. Testes less than 50% of full size
F4	Ovary contains opaque oocytes, clearly visible when ovary is cut	M4	Testes filling but no sperm visible when cut
F5	Ovary contains hydrated cells but does not run	M5	Sperm visible when cut but testes do not run
F6	Like stage 5 but runs under moderate pressure	M6	Testes run under moderate pressure
F7	Ovary slack, contains slime and sometimes remaining oocytes	M7	Testes can be quite small, fluid remaining in sperm duct

Adapted from Bromley (2000) and (Tomkiewicz et al., 2003).

Table 2. Tally of all pair-wise comparisons between assessors by maturity stage for the combined plaice samples. Percentages of raw agreement are given.

Plaice	M1	M3	M4	F1	F2	F4	Agreement
M1	105	13	1	19	3	0	74%
M3		126	49	10	3	0	63%
M4			150	1	3	0	74%
F1				74	83	0	40%
F2					141	10	58%
F4						169	94%