INVESTIGATIONS INTO THE QUALITIES OF FARMED, FRESH SOUTHERN BLUEFIN TUNA, AIR-FREIGHTED FROM PORT LINCOLN, SOUTH AUSTRALIA TO TOKYO, JAPAN

ALISTAIR EWAN DOUGLAS BA-BSc (Hons)

SCHOOL OF BIOLOGICAL SCIENCES, FACULTY OF SCIENCE AND ENGINEERING, FLINDERS UNIVERSITY, ADELAIDE, AUSTRALIA

March 25, 2007

TABLE OF CONTENTS

DECLARATION

I, Alistair Ewan Douglas, certify that this thesis does not incorporate without acknowledgment any material previously submitted for a degree or diploma in any university; and that to the best of my knowledge and belief it does not contain any material previously published or written by another person except where due reference is made in the text.

………………………………………………………….

Alistair Ewan Douglas

ACKNOWLEDGEMENTS

The challenges involved in multi-disciplinary and multi-industry research in two nations were significant. For example, directly involved in one experiment were a tuna farm and its staff (Blaslov FIshing), a processor (Australian Bight), a freight forwarder (Danzas), five Australian scientists, students and staff (Flinders University and Allan Bremner & Associates), six Japanese scientists, students and staff (Tokyo University of Fisheries), five staff of an auctioneering company (Toichi), four staff of a wholesaler (Sugahei), and finally, eight research staff and forty five general staff of a major Japanese trading company (Nippon Suisan). Organization of this particular experiment was not only challenging – its execution was complicated by a storm and the tuna involved delayed in Sydney for twenty four hours due to the dangers of the originally booked aircraft breaking apart during landing (as lighter flowers had been previously loaded onto the plane unbalancing the trimmings of the aircraft). The times of just being allowed on a local plane covered in tuna blood, crossing the equator, and not sleeping on a bed (or futon) for over sixty hours to do an experiment really did make the beer and sake taste that much better at the end of the day(s)!

This work therefore resulted from the collaborative efforts, both directly and indirectly, of so many people representing research, government and commercial groups in Australia and Japan, that it would require too much paper here to personally thank everyone. However I must name the following:

In the Australian Industry: Justin Nelligan & Ric Lehmann of Blaslov Fishing; Rick Kolega, Semi Skoljarev & David Warland of Sekol; Stephen Moriarty of Southern Waters; Terry Richardson of Australian Bight; Brian Jeffriess of TBOAA., and David Ellis of Fishing Industry Services. In Australian R&D organizations: Peter Montague & Emily Downes of Aquafin CRC; Stephen Clarke of SARDI; Stephen Nottingham of CFT; Mark Thomas, Kelie Stead, staff and resident researchers of the LMSC.

In the Japanese industry: Tatsumi Taguchi, Ryozo Moriwaka, Masuo Murakami & Kenichi Kurihara of Toichi; Tsunenori Iida, Toichiro Iida & staff of Hicho; Takeshi Ueda of Naito; H. Tanaka & staff of Sugahei; Joji Kojima & staff of Yamajo. In the Japanese R&D organizations: Takafumi Arimoto, Toshiaki Ohshima, Etsuo Watanabe, Chiaki Imada, Naoko Hamada, Takeshi Kobayashi, Sohn Jeong Ho, staff and students of Tokyo University of Fisheries; Testu Mori, Tomoko Kohata, Itaru Shioya, Kazuyuki Yuguchi, & Minako Obama of the Central Research Laboratories of Nippon Suisan

In closing I would like to thank my supervisors Dr. John F. Carragher and Dr. Philip M. Thomas for all your efforts and advice along the way. I am greatly indebted to Prof. H. Allan Bremner - this research would not have been completed without your involvement, your support, and timely and detailed correspondence – thank you. Finally, my family and friends for your unconditional support throughout – Cheers!

David this is a little bit better mate…

ABSTRACT

The establishment of the Aquafin CRC enabled Japan-based research into the qualities of farmed, fresh Southern bluefin tuna to be conducted. This occurred via industry collaboration in both Australia and Japan, and through the establishment of a memorandum of understanding between Flinders University of South Australia and the Tokyo University of Fisheries (now the Tokyo University of Marine Science and Technology), and an agreement between the Aquafin CRC and Nippon Suisan, the product was profiled, and instrumental and sensory investigations into the qualities of this valuable product were able to be developed and undertaken.

Although the three major cuts of tuna white muscle, known as Akami, Chutoro, and Otoro, are compositionally different, they were shown to have similar patterns of change post-mortem for a selection of bio-chemical parameters commonly associated with ‘quality’, potentially allowing for the indirect assessment of the more valuable cut (Otoro) from the destructive sampling of the less valuable cut (Akami).

Further, the establishment of a correlation between expert subjective assigned ranks of ‘quality’ and a ratio of derived red, green, and blue (RGB) values from digital images of the flesh, offers a new objective quality assessment technique that is both rapid and non-destructive. In addition, a balanced and statistically robust analytical protocol was developed for the sensory assessment of the whole carcass qualities of tuna flesh. The protocol allows for the affect of any

on-farm or in-chain manipulations on the sensory properties of the flesh that are directly perceptible to consumers to be assessed.

As the product has a reputation for short colour shelf-life on the market, the effects of using vitamin supplements as a counter measure (as per the industry practice) on the concentrations of vitamins in the flesh and on the colour shelf-life of the end product were investigated. Vitamin supplementation was categorically proven to aid in the colour retention of the flesh of farmed Southern bluefin tuna with low to medium levels of fat both in Australia and in Japan.

Harvest stress is known to affect the qualities of fish flesh, and in this study the effects of a prevalent industry harvesting practice on a selection of sensory and biochemical quality related characteristics of tuna flesh were investigated. Although there were no significant differences in the majority of the sensory and biochemical indicators of quality between fish harvested at the beginning or at the end of a commercial tuna harvest, expert-calibrated RGB ratios and the sensory descriptors of transparency and brightness resulted in significant deleterious effects of harvest stress on the Akami and the valuable Otoro sections respectively.

Finally, the time-temperature management of chilled tuna carcasses when air-freighted to Japan, as well as the effects of shipping on the day of harvest or the day after harvest on flesh quality were investigated. Within the cold chain, the most likely periods when temperature control could be violated were shown to be during the loading and off-loading of the tuna coffins at the airports. And, although there were no statistical significant differences between the sensory and biochemical parameters measured from fish shipped on the same day as harvest when compared to those shipped a day after harvest, averages favoured the latter where recorded carcass temperatures were lower and more stable.

Finally this body of work demonstrated that collaborative market-based research can be undertaken in Japan, and that product quality needs to be measured in a way that is sympathetic to customer culture and expectations.

1. INTRODUCTION

1.1 Background

Southern bluefin tuna (SBT) – *Thunnus maccoyii* – is a member of the Thunnus family of tunas which include Northern bluefin tuna, Yellowfin tuna and Bigeye tuna. They are large, fast swimming pelagic fish found throughout the Southern Hemisphere mainly in the waters between 30 and 50 degrees south. Their only known breeding ground is in the Indian Ocean south east of Java, from where the juveniles migrate down the coast of Western Australia, around Cape Leeuwin to school in the waters of the Great Australian Bight (Caton, 1991; Fig. 1). It is in these waters that the Southern bluefin tuna fishing and canning industry of Port Lincoln in South Australia was established in the 1950s by a group of pioneering immigrants (Fig. 1.1).

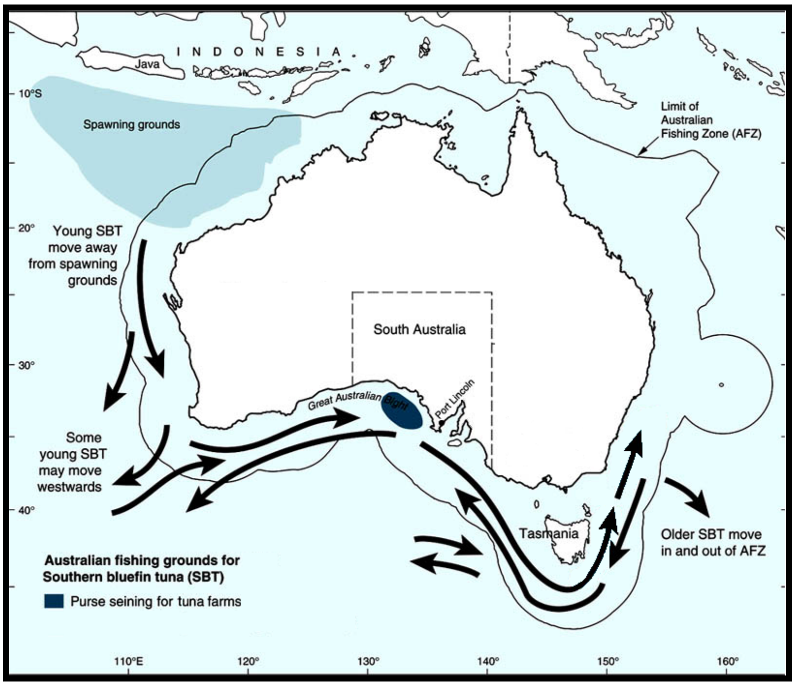
INDONESIA
Java
10°S
Spawning grounds
Limit of Australian Fishing Zone (AFZ)
20°
Young SBT move away from spawning grounds
South Australia
Great Australian Bight
Port Lincoln
30°
Some young SBT may move westwards
40°
Tasmania
Older SBT move in and out of AFZ
Australian fishing grounds for Southern bluefin tuna (SBT)
■ Purse seining for tuna farms
110°E
120°
130°
140°
150°
160°

Figure 1.1: Spawning, fishing grounds, and migratory routes around Australia for Southern bluefin tuna (*Thunnus maccoyii*). Source image: Caton (1991).

The species however, was over fished in the 1970s and 80s, and in response to poor catches and poor economic returns from the canned tuna market, a research and development venture involving the Tuna Boat Owners Association of Australia, the Japanese Overseas Fisheries Cooperation Foundation and the South Australian government was established in the early 1990s to examine the feasibility of transporting live, wild captured fish from the Great Australian Bight to the waters off Port Lincoln for shipment to the lucrative Japanese sashimi market. The relative success of this first trial led to a two year program funded by the Fisheries Research and Development Corporation (FRDC) to investigate the holding, maintaining and marketing of the product (Clarke & Bushell, 2001). It is from these beginnings that the SBT farming industry of Port Lincoln evolved.

Instead of being landed at sea in December to March each year, the industry now corrals the tuna in purse seine nets and transfers them underwater into tow cages. These cages are then towed from the fishing grounds to the waters of the Spencer Gulf off Port Lincoln where they are once again transferred into larger grow out cages. The tuna are kept between two to six months in the Spencer Gulf being fed a variety of baitfish species. When they are ready for market they are hand-harvested, placed immediately into iced sea water, processed on land, either placed into freezer containers or loaded into refrigerated trucks, and then shipped or air-freighted to Japan respectively.

A Commonwealth Government agreement in 2001 saw the establishment of an aquaculture focussed Collaborative Research Centre (Aquafin CRC) with the aim of providing critical technologies for the rapid and sustainable growth of finfish (particularly Atlantic salmon (*Salmo salar*) and SBT) aquaculture in Australia. Set up for a seven year period (2001-2008) with a planned investment into Australian aquaculture totalling $34 million, the Aquafin CRC provided the Farmed Southern Bluefin Tuna Aquaculture Sub-Program, initiated by the FRDC in 1997, with a substantial funding boost.

Within the R&D sub-program there are five major areas of research including farm husbandry and management, feeds and nutrition, environmental monitoring and mitigation, fish health, and product quality. In the early years, the industry members and researchers of the product quality team had a limited understanding of the raw product they were producing and shipping to Japan and what quality characteristics constituted a sashimi grade product. Therefore, it was first necessary to characterise the product qualitatively into its parts, and quantitatively via physico-chemical analyses, before determining if there were effects of any pre- and post-harvest processes on the qualities of the product. Along with a variety of instrumental techniques, a subjective flesh colour ranking scheme was developed to investigate and monitor the effects of various treatments such as harvest stress and vitamin supplementation on the colour stability of the product. This sensory method soon became a cornerstone of the research program and a well trained and experienced panel now exists at the Lincoln Marine Science Centre in Port Lincoln.

The funding boost provided by the Aquafin CRC brought with it opportunities to investigate the qualities of the product in its market in Japan. To facilitate this, a Memorandum of Understanding was signed between CRC participant Flinders University and the Tokyo University of Fisheries (now the Tokyo University of Marine Science and Technology). This allowed for research to be undertaken not only in Australia but also Japan, and for analytical techniques to be shared, developed, and applied in an atmosphere of collaboration. Another major step was the signing of a collaborative agreement between the Central Research Laboratories of the large Japanese seafood importer and processor Nippon Suisan Pty. Ltd. and the Aquafin CRC. These agreements made possible physico-chemical and sensory research into the qualities of Australian farmed Southern bluefin tuna at the point where it is sold and consumed. The following research chapters detail the methods employed, developed, and investigated in both countries to define, sample, characterise and measure the instrumental and sensory qualities of this unique product.

1.2 Defining Quality

> *"Quality is an unusually slippery concept - easy to visualise and yet exasperatingly difficult to define"* Garvin D.A. (1988) Managing Quality: The Strategic & Competitive Edge, The Free Press, MacMillan Inc. New York.

According to Payson (1994), prior to any discussion on quality, a specific perspective must be established with regard to its meaning. The term quality is positioned third within the fourteen Aristotelian categories, following substance and quantity, and its various definitions occupy more than two pages of the Oxford English Dictionary (OED 2nd Edition, 1989). The word has been used to refer to the character, disposition, nature, capacity, skill, accomplishments, title, social position, profession, fraternity, and the mental and moral attributes of both humans and animals. It is used to define objects by their attribute, property, manner, style, habit, power, substance, nature, and kind. Its synonyms are many. Furthermore, quality is contextual and relative. Thus, the quality of identical items can be judged differently at either the same time in a different context or in the same context at a different time (Meiselman, 2001).

In developed economic societies the importance of product quality to both producers and consumers is rarely questioned, however, the determinants of quality and its meanings are often poorly, if at all, defined (Bremner, 2000; Meiselman, 2001). Garvin (1988) asks the questions: Is quality objective or subjective? Is it relative or absolute? Is it timeless or socially determined? Can it be divided into narrower and more meaningful categories? According to Payson (1994), as modern humans are economic beings, quality legitimizes us as providers of goods and services, and, in some sense is our raison d'être.

Economists study quality and ways to measure it in order to figure out quantity. This is opposed to businesses, political organizations, and research organizations whose struggle is to foster its improvement (Payson, 1994). The same author proposes that a "good's quality is an inherent aspect of the good itself, whether or not one can actually measure it". However, such a definition only leads to the frustrated retort "*I know quality when I see it!*" - a statement often used by middle management to their subordinates when struggling to define quality despite it being the goal of all firms (Taormina, 2001).

A global institutional approach to defining and standardizing the qualities of both products and services, the International Organization for Standardization (ISO), claims that a lack of standardization can affect the quality of life itself. That the standardization of screw threads helps to keep chairs, children's bicycles and aircraft together, and that for the disabled, for example, they are able to access and use consumer products, public transport, and buildings because the dimensions of wheel-chairs and entrances are standardized.

In his attempt to define quality, Garvin (1988) proposed the following five categories; Transcendent Quality, Product-Based Quality, User-Based Quality, Manufacturing-Based Quality, and Value-Based Quality. Transcendent Quality is synonymous with innate excellence and somewhat beyond definition but attainable via experience. Product-Based Quality is viewed as a precise and measurable attribute of a product. User-Based Quality centres on the premise that beauty lies in the eyes of the beholder. Manufacturing-Based Quality is based on the conformance to requirements in engineering and manufacturing practices. And finally, Value-Based Quality provides conformance or performance at an acceptable cost or price. However, a much more simple approach to quality was proposed by Crosby (1979) where he states that *"Quality means conformance to requirements, and that's all it means. If you start confusing quality with elegance, brightness, dignity, love, or something else, you will find that it has different ideas. Don't talk about poor quality or high quality. Talk about conformance and non-conformance."* Although suitable for screws, such a definition alone would hardly be welcomed by the food and beverage industries other than the technologists.

The Total Food Quality Model (TFQM), originally proposed by Grunert et al. (1996), is an attempt to integrate a number of approaches to analyse consumer quality perception and decision-making. The model distinguishes between before and after purchase evaluations with the dimensions of quality categorized into search, experience and credence characteristics (Darby & Karni, 1973). Which category a product is placed in depends on when the consumer can ascertain quality attributes. A search quality (like the appearance of a piece of meat) can be evaluated before the purchase, an experience quality (like the taste of the meat) can first be evaluated after the purchase, and a credence quality (like the healthiness of the meat) can, under normal circumstances, not be evaluated by the average consumer at all, but is a question of faith and trust in the information provided. Many characteristics of a food product, like taste, cannot be ascertained before purchase. Therefore most food products have search characteristics to a limited degree. In order to make a choice, the consumer will develop expectations about quality — but it is only during and after consumption that experience quality can be determined, and even this is limited in the case of credence characteristics like the healthiness of a product. The consumer and technologist alike must therefore rely on attributes that, in their experience, imaginary or not, link with the experience quality

The term quality, according to Bremner (2000), is properly used in advertising, sales, and marketing to create an impression in the minds of subjects without having to be specific about the meaning. However, in the scientific field a more specific meaning of the word quality is required, and along with the words fresh or freshness, the author claims it is probably the most misused word in food science. Further, according to Meiselman (2001), despite food quality being multidimensional, most attempts to define and measure it in the food sciences are one dimensional such as the sensory characteristics or the microbiological status of food. Studies in food quality have centred mostly in product development and have utilized technical approaches to food science and technology, microbiology, consumer research, and market research (Meiselman, 2001). However, definitions of quality may vary according to these differing viewpoints. A technologist may list only safety, nutrition, availability, convenience, integrity, and freshness as the quality defining attributes of foods. Other qualities such as value for money, legal value, technological value, socio-ecological value, psychological value, and political value could be included in a definition when the viewpoints of other stakeholders and professionals are considered (Bremner, 2000).

With quality a national obsession in Japan, and the colour, shape and arrangement of a meal as carefully thought out as a painting (Garvin, 1988; Yoshida & Sesoko, 1989), how does the Japanese market define the quality of tuna used in sashimi or sushi – the origins of which dates back over 1200 years (Tayama, 1981)? Do Northern bluefin tuna, caught in the waters off Japan since the Manyo Period (8th century) require a separate definition to that of a Southern bluefin tuna, which Japanese fisherman only commenced fishing for in 1952-53 in the southern hemisphere when on-board freezer technologies were developed? Are definitions universal and static or do they change according to the time of season, fishing ground, region, or fish size? Can the quality of an individual tuna be assessed or is the quality of each of the differing cuts of a tuna assessed separately? Do the quality definitions of these differing cuts change depending on form or intended use – raw cuts used in sashimi or sushi, minced in sushi, or baked?

Finally, is the assessment of quality standardised or do the definitions change between the differing stakeholders – the wholesalers, retailers, restaurateurs, chefs, consumers, regulators, etc.? Amongst consumers, do notions of quality change depending on whether sea birds or dolphins were caught along with the tuna, whether the fishery is sustainable, whether the species is rare and expensive, or will it vary whether the consumer is a man or a woman, or between women who are or who are not pregnant etc.?

Tunas, according to Ebisawa (1996), were not considered a high quality sushi ingredient until the beginning of the Second World War. Indeed, one of the first specialist sushi restaurants established in 1684 stated on the entrance curtain that '*As tuna goes off quickly it is not served here*'. At the beginning of the Showa Period (1926-1989) the high fat *toro* regions of the tuna became regarded as higher quality than the low fat *akami* regions. The reason behind the fascination with *toro*, Ebisawa (1996) claims, is deeply tied to the 'violent surge' of American culture following the Second World War and the westernization and 'fattening' of the food culture of the Japanese.

As a result of this, the high fat *Hon Maguro* (Northern bluefin) and *Minami* or *Indo Maguro* (Southern bluefin) are considered the best tuna species with a single piece of sushi (approximately 10g) costing as much as 2-3000 yen (Tayama, 1981; Ebisawa, 1996). Southern bluefin however, according to Tayama (1981), is a relative newcomer and Japanese consumers have not been acquainted with this species for very long when compared to Northern bluefin. According to Ueda (2003), a 'good' Southern bluefin will be comparable to a Northern bluefin in flavour, and thus, they are considered number two (for raw consumption in Japan).

The wholesale market for tuna in Japan is artisanal with some wholesaling operations being family concerns spanning ten or more generations. There is neither a definitive nor market-wide systematic grading scheme used to describe and assess tuna quality. Instead, it is referred to and communicated colloquially amongst wholesalers who are trained over many years to distinguish 'good' from 'bad' quality tuna. This is done prior to auction, where wholesalers assess tuna quality visually and tactilely, and descriptors used are based mostly on the colour, fat, and moisture levels of the flesh, and the shape of the tuna. Individual wholesalers will occasionally use symbols to represent the quality of the tuna they are interested in bidding on in the auction and will write them alongside an individual fish's auction number. Such symbols can be developed by individual wholesalers for their own use or handed down and shared amongst other employees of the same company.

There are many general terms used to describe the quality and shape of tunas on the market. Terms such as the onomatopoeic '*gari'* refers to the sound a knife makes when inserted into the carcass of a low fat specimen, and 'rakkyo', which refers to a fish that's shape resembles that of a shallot where the head appears large and the tail is thin. These are often fish that may have recently spawned and have minimal fat in their flesh and are therefore considered 'low quality' in the current market. Shallots (*rakkyo*) and ginger (also *gari* in Japanese) are often used as condiments when eating raw tuna – possibly along with these low quality tuna as they may be considered necessary to add flavour (Tayama, 1981). Other examples include *mizumaguro*, or 'water tuna', which are tuna whose flesh lacks consistency and is too moist when eaten, and *akabero,* which refers to flesh that has a red jelly like appearance and texture (Tayama, 1981). In addition, a whole range of terms and expressions exist to describe the flesh of diseased, damaged, or poorly processed tuna (for a full list see pages 15-17 of *The Australian Tuna Handling Manual – A Practical Guide to Industry* (Erica Starling & Geoff Diver). Seafood Services Australia, Queensland, Australia.

As a large quantity of tuna is sold on auction floors throughout Japan, it may be argued that the average price paid per kilogram by Japanese wholesalers is a useful determiner of ‘poor’ and ‘good’ quality tuna. However, for similar reasons economists have problems pricing oil, gold or water, tuna catch and supply can vary greatly on any given day or within any given season. Furthermore, and unlike those other commodities, demand for tuna is elastic as there are substitutes on the market – both as a source of sashimi tuna and as a source of protein. Moreover, buyer and consumer behaviour can be erratic with auction battles leading to extreme prices such as the 20.2 million yen (AUD$310,000) paid for a single 202kg Northern bluefin tuna by a wholesaler in Tokyo who told reporters after the auction, "I just wanted to buy the highest quality tuna" (Anon, The Japan Times, 2001). Although such outliers can be removed from analyses, the issue of arbitration and the need to collect a whole range of market-related data (e.g. indicators of demand, supply, price and availability of substitutes etc.) would complicate the use of auction price as a meaningful quality indicator to assess the outcome of flesh quality experiments over time.

With no definition and standardised grading system on the markets in Japan, tuna quality as it is assessed and communicated appears to fit into the first of Garvin’s (1988) five definitions of quality (transcendent quality) where the author claims that quality cannot be defined precisely - that it is an un-analysable property one learns to recognise only through experience. Such a definition and system of assessment, however, is not particularly suitable for scientific investigation. Therefore, it is necessary to remove the psycho-socio-economic aspects of quality, and examine tuna solely as a muscle food that is comprised of water, proteins, fats, carbohydrates, vitamins and minerals, and a variety of other organic and inorganic constituents; with the relative mix of these constituents determining the conformation, appearance, odour, texture, flavour, nutrition, and safety of the product at any one time prior, during, and after consumption. These are the physico-chemical and sensory qualities of a single cut of tuna muscle - its ‘qualitas’. It is only when these qualities are combined with hedonic properties, such as the ‘peace of mind’ a consumer may feel if the tuna came from a sustainable fishery or the feeling of ‘exclusivity’ a consumer may feel if consuming a rare tuna etc., that the ‘total quality event’ peculiar to the individual consumer and the cut of tuna in question is produced.

In the case of fisheries and aquaculture, as with other industries, it is the role of marketers, industry organizations, technologists, and company management to identify, define and improve the psycho-socio-economic qualities of a product and service where it is possible to do so. Alternatively, it is the role of fishers or farm managers, processors, and distributors to ensure the product meets defined safety standards and possesses the physico-chemical and sensory qualities that best satisfy its end users. It is this concept that best reflects the ISO 8402 Standard for Quality which defines quality as "the totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs". These ‘needs’ can be identified by establishing links between the physico-chemical and sensory qualities of the product to the results of blind consumer preference testing. Prior to accomplishing this, however, it is first necessary to identify and/or develop methods to directly measure and profile the inherent variations in both the physico-chemical and sensorial qualities of the product.

1.3 Measuring the Qualities of Fish

The quality attributes of fish can be analysed instrumentally or sensorially. However, as fish is a food product the ultimate judge of quality is the consumer and instrumental methods need calibrating to sensory techniques (Gill, 1995). The appeal of instrumental techniques, according to that author, is that they allow for the setting of quantitative standards such as the tolerance levels of chemical spoilage indicators, and eliminate the need to base decisions on personal opinions and time-consuming microbiological methods.

1.3.1 Instrumental Techniques

There are many instrumental techniques available for making an assessment of the qualities of fish – be they measures of body composition or freshness. Body composition assessment methods can be divided into five levels: the atomic level, the molecular level, the cell level, the tissue level, and the whole-body level (Durnin, 1995; Duerenberg & Schutz, 1995). All levels are related and one can calculate total body composition from each level, assuming constant and equal relationships in all individuals. Methods can be either direct (chemical analysis), indirect (making use of data based on chemical analysis), or doubly indirect (based on a statistical relationship between easily measurable body parameters and data obtained by direct or indirect methods) (Duerenberg & Schutz, 1995). Nearly all of the techniques used for estimating body composition are indirect measurements. That is, they measure some physical property of the body which is related to body composition, and then make use of the assumed constancy of the relationship to calculate composition (Nord & Payne, 1995). In Neutron Activation Analysis (NAA), for example, the body is infiltrated with fast neutrons of a known energy level. The neutrons are captured in the body by specific chemical elements, depending on their energy, resulting in the formation of specific isotopes with initially higher energy levels. The energy is then emitted in the form of gamma rays with a well-defined energy level dependent on the isotope formed. It is from these energy emissions that the amounts of nitrogen, calcium, chlorine, sodium, phosphorus, and oxygen can be calculated, and then, via stoichiometric relationships, it is then possible to determine the amount of body proteins, bone mass, extracellular water, and fat levels (Duerenberg, & Schutz, 1995).

The focus of body composition analysis in the aquaculture industry is often on the lipid and protein fractions as they relate to and affect fish growth and reproduction, as well as the storage, appearance, and organoleptic properties of the end product. This is especially the case for tuna where the fatty part of a carcass is the most valuable cut and its levels can have a significant effect on its colour, texture, and taste (Rye, 1991; Sigurgisladottir *et al.,* 1997). Water, as a key component in food systems including fish flesh, is also a fraction of interest as it influences most process variables, product characteristics, and stability attributes. Moisture levels influence all diffusion-controlled reactions (e.g.: enzymatic activity, crystallisation processes, and browning) and usually the most important factor determining the thermodynamic properties of flesh, and therefore, its quantification, along with fat and protein levels, allow for a better understanding of flesh qualities (Jepsen *et al.*, 1999).

With the freshness of fish being a major quality, health and safety concern, a number of instrumental techniques to gauge fish freshness have been developed. The techniques can be physical or chemical and can be linked to the sensory qualities or microbiology of the flesh under certain storage conditions. The most common methods measure the accumulation and/or degradation of volatile compounds, oxidised lipids, or adenosine tri-phosphate, or changes in the texture, microstructure, or electrical properties of the flesh (Gill, 1995). Instrumental measures of freshness are often used in the resolving of issues regarding the marginal qualities of fish. However, unlike sensory techniques, they can be time consuming, lack sensitivity, and unable to determine notions of ‘good’ and ‘bad’ quality (Gill, 1995).

1.3.2 Sensory Techniques

The assessment of fish ‘quality’ was for centuries exclusively a sensory appraisal of the aesthetic appearance and freshness of fish (Nielsen, 1995). Although still a satisfactory method for fishmongers purchasing from artisanal fishers whom after fishing for a few hours return to sell their catch while the fish is still alive or very fresh (Huss, 1995), it is not satisfactory for the now global fish market where fishing grounds and consumers are often thousands of miles apart. In order for regulators, wholesalers and retailers to be sure that fish distribution channels are supplying consumers with safe and healthy fish and fish products it is necessary that sensory methods be performed scientifically under carefully controlled conditions so that the effects of the test environment, personal bias, etc., can be reduced (Nielsen, 1995).

More objective approaches to the sensory assessment of fish freshness included an indexing system developed by the Torry Research Station in the 1950s, and the EU scheme, introduced in 1976, which grades fish into one of three quality levels - E (Extra), A, and B where E is the highest quality and anything below B is the level where fish should be discarded for human consumption (Neilsen, 1995).

The Quality Index Method (QIM), originally developed by the CSIRO Tasmanian Food Research unit is now widely used throughout the EU (Hyldig & Green-Petersen, 2004). Based on characteristic, well-defined changes in several significant sensory parameters, QIM is a practical rating system of the freshness of raw fish species. Inspecting fish for changes in their outer appearance, a score from 0 to 3 demerit index points is assigned to each characteristic (eyes, skin, gills, odor etc.). A score of zero is given for very fresh fish while increasingly larger scores result as a characteristic deteriorates (Bremner *et al.,* 1987; Neilsen, 1995). The scores for all characteristics are then summed to give an overall sensory score of an individual fish. Further, as the QIM scale has been devised to produce a linear correlation between the demerit score and storage life on ice, it is possible to predict remaining storage life on ice.

With limited control over the compositional qualities of fish caught in the nets or on the hooks of fishers, instrumental and sensory techniques used to assess fish quality have focussed on measuring the so-called freshness of fish in order to identify ways to best preserve quality through post-harvest approaches. Aquaculturalists though, have greater control over the compositional qualities of the fish they harvest and therefore their concerns have expanded from not only measuring and checking the marginal qualities of fish, such as its fitness for consumption, to measuring the hedonic qualities of the product – those attributes and characteristics that make a fish look and taste ‘good’. Sensory evaluation, therefore, is now of great interest to the aquaculture industry as a tool to scientifically measure the effects of on-farm or in-chain practices on the characteristics and attributes directly related to the 'eating experience' of their customers when consuming the end product. However, as sensory evaluation techniques are mostly used in the assessment of processed foods, it is necessary to review the processes and considerations of sensory evaluation and assess the unique challenges of applying a particular technique to an unprocessed fish product such as farmed Southern bluefin tuna.

1.3.3 Process of Sensory Evaluation

The concept of sensory evaluation began to appear in the literature post World War II, and although a relatively new discipline, along with microbiological safety and nutrition, the sensory properties of foods are one of the primary determinants of our food preferences (Frijters, 1984; White, 1996). The modern discipline of sensory science draws upon the theories and practices of food science, physiology, psychology, and statistics to arrive at analysable responses from stimuli perceived by the five senses (Piggott et al., 1998). According to Frijters (1984), there exist three main elements (the stimulus object, the sensory perception, and the sensory response), and two relationships (psychophysical and psychometrical) that are of keen interest to the sensory scientist. Psychophysics deals with the relationships between the physical properties of the stimulus object and the sensory characteristics of perception. Psychometrics confronts the relationships between the sensory stimuli perceived and the responses of subjects (Frijters, 1984).

The methods that examine the psychometric and the psychophysical relationships between foods and humans can be considered either affective (subjective), and involve the examination of consumer preferences and/or their acceptance of products, or analytical (objective) using trained panellists and which centre on the measurement of the qualities of foods and their differences or similarities (Larmond, 1987). Complicating the study of these relationships is the fact that sensory perceptions are private events and therefore, as they are not directly observable, sensory measurements are classified as derived measures (Frijters, 1984). Consumer trials using untrained subjects can reveal a particular demographics' acceptance, preference, or bias toward a particular product or product's characteristic, but they cannot quantitatively describe or discern the sensory characteristics of, or between, two products with acceptable degrees of accuracy or precision. To achieve the latter we use sensory evaluation techniques with panellists whose sensitivities have been tested and who have been trained to differentiate, describe and evaluate the characteristics of a particular subject matter. Although sometimes criticised for its lack of reliability due to subject error, sensory evaluation of the properties of foods is considered to be the most direct and the most sensitive measurement technique when compared to quantitative methods (Frijters, 1984).

The first step in the development of a sensory evaluation strategy of a food or material is the identification of the test objective (Stone & Sidel, 1987). Once identified the selection of an appropriate testing method will depend upon the product and its requirements, the logistical and cost constraints of the chosen test, and the qualifications and availability of test subjects. Following test selection it is necessary to consider the measurement technique, the type of response scale, the experimental design, and how the response data is to be analysed (Fig. 1.2).

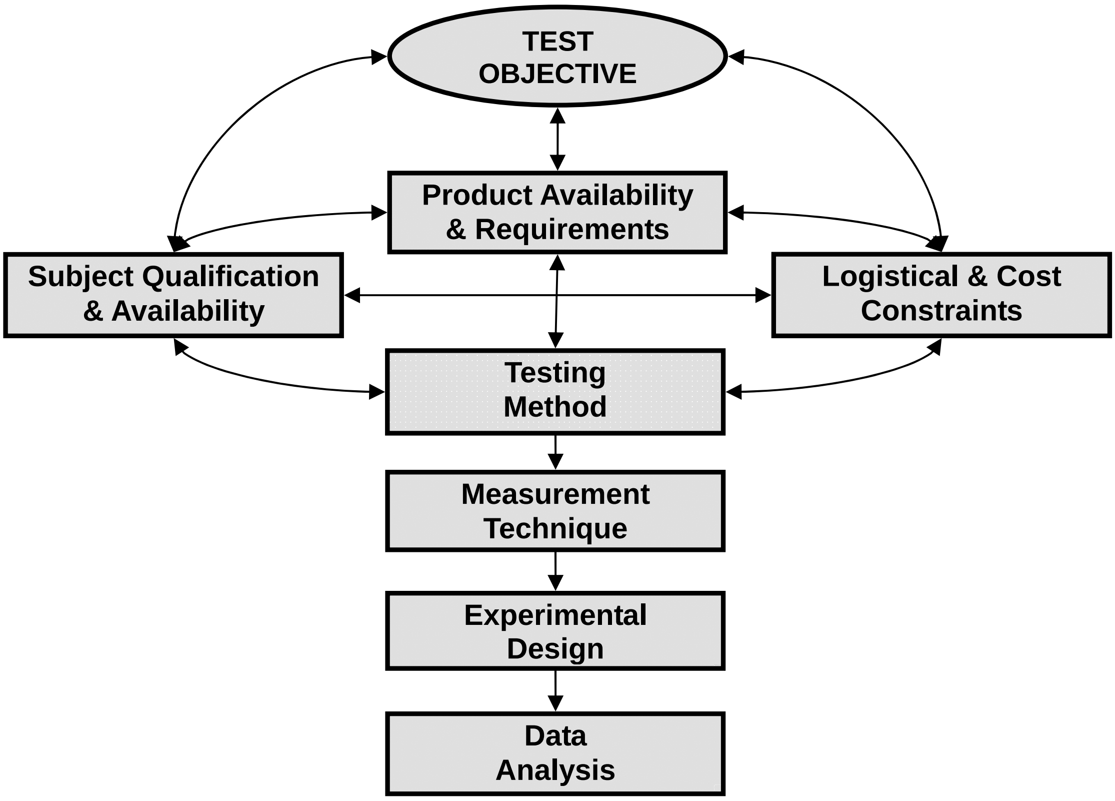
TEST OBJECTIVE
Product Availability & Requirements
Subject Qualification & Availability
Logistical & Cost Constraints
Testing Method
Measurement Technique
Experimental Design
Data Analysis

Figure 1.2: Connectivity and order of the considerations in the formulation of a sensory testing strategy.

Identifying the Test Objective

Central to the success of any sensory test is a clear understanding and statement of the test objective (Larmond, 1987). The stated test objective, the testing method, the experimental design and analysis, and measurement technique all need to tackle the identified problem and relate well to one another in order to yield reliable and valid product information (Stone & Sidel, 1993).

Testing Methods

Test objectives in sensory science fall into one of two major forms of sensory testing method – difference or descriptive tests (Piggott, 1998). Difference techniques aim to identify whether or not products are perceivably different, and descriptive methods attempt to identify and measure the sensory composition of foods, or determine the presence and/or intensity of particular attributes of a food (Piggott, 1998).

Difference tests are procedures used in many disciplines including biology, psychology, economics and market research to measure comparative judgements and choices (Frijters, 1984). In sensory science they are used to discriminate between two different types of stimuli or products. The simplest difference test is called a *duo test* where two products are presented to one or more subjects with the instruction to select the stronger of the two with regards to a pre-specified attribute. A *duo-trio test* removes the need of attribute specification prior to the test, as a sample of one of the two products to be examined is first used as a reference. The subject is then asked to select which of the two products differs most from the reference (Frijters, 1984).

Another type of difference test, known as a *triangle test*, presents subjects with multiple sets of three coded samples with half the sets containing two samples of product A and one sample of product B, and the other half containing two samples of product B and one of product A. The subjects are then informed that two of the samples are the same and one is different, and for the test, instructed to identify the odd sample of the three (Frijters, 1984; Larmond, 1987). In the instance where the assessor cannot detect a difference but is forced to choose the odd sample out, *triangle tests* are more efficient than *duo-trio tests* in that the probability of selecting the correct sample by chance is 33% rather than 50%. However, with less tasting required, *duo-trio* tests can be preferred to triangle tests when strongly favoured products are being investigated (Larmond, 1987). Variations of the same theme include the *tetrad test* and the *hexagonal test* where multiple reference and treatment samples are provided.

Ranking tests are difference tests which present subjects with three or more samples and subjects are instructed to order the samples according to the levels of intensity of a particular characteristic. Although rapid, the test provides no indication of the size of the difference between samples and as samples are evaluated in relation to each other results from one set of ranks are not comparable to the results from a differing set of ranks (Larmond, 1987). According to Stone & Sidel, (1993) it is probably this latter limitation that has resulted in the infrequent use of ranking in sensory evaluation.

Difference tests provide no indication of the dimension of difference and, as the sensory nature of ‘oddity’ is often not specified, the subject is left to him/herself to decide upon the sensory attributes which are relevant for discrimination (Frijters, 1984; Larmond, 1987). Mostly used in quality control or ingredient-substitution investigations, where the products are relatively homogenous, all sensory difference tests are forced-choice procedures in which the subject is required to select a sample that is different even though the subject may not be able to detect any discernible differences (Frijters, 1984; Larmond, 1987). When not discernible, the choices, considered random guesses, are either correct or incorrect and the probability of a correct response is easily determined with the use of binomial or chi-square testing where it is possible to calculate whether the differences in the responses were due to chance (sampling variability) (Frijters, 1984). As sensory difference tests are based on a statistical comparison of the distribution of correct and incorrect responses and the expected theoretical distribution of random responses, they are not founded on the principles of sensory perception according to Frijters (1984), but on the combined principles of guessing behaviour and probability theory (Frijters, 1984).

Descriptive sensory analyses utilize trained and experienced subjects that examine and evaluate food products to provide detailed descriptions of appearance, flavour, and texture. Of available descriptive methods flavour profile analysis (FPA), texture profile analysis (TPA) and quantitative descriptive analysis (QDA) are the most widely known and used (Larmond, 1987). These descriptive scaling methods attempt to describe the perceptible factors, the intensity of each, their order of perception, aftertaste, the overall impression, and qualitatively and quantitatively describe the mechanical and geometric characteristics of foods (Larmond, 1987). QDA combines descriptive analysis, unstructured scales, and repeated measures to characterise the sensory attributes of products in order of appearance, the intensities of each attribute, and then statistical techniques to determine whether or not significant differences exist between the intensities of sensory characteristics (Larmond, 1987). Although highly valuable tools these descriptive sensory techniques require highly trained and motivated subjects (Larmond, 1987).

Magnitude estimation is an experimental technique that attempts to quantitatively scale how much of a particular sensation subjects are experiencing. Subjects are presented with a series of samples that vary in a particular characteristic and are instructed to assign a number to the first sample (or a number is assigned by the experimenter), and then rate each following sample in relation to the first. If, for example, a subject rated the sweetness of a liquid first in a series as being ‘10’ then any of the following samples considered ‘half as sweet’ by that subject would score a 5, and a sample considered ‘twice as sweet’ would score a 20 (Snodgrass *et al.,* 1985; Larmond, 1987).

In summary, *triangle, duo*, and *duo-trio tests* indicate a difference only between two samples, *ranking tests* indicate if the samples differ in a particular characteristic and the direction of the difference, descriptive scaling methods provide information on the size and direction, and magnitude estimation provides information on the proportions of differences (Table 1.1; Larmond, 1987).

Test Name	Type of Test	Information Obtained			
		Difference	Direction	Size	Proportion
Duo, *Duo-Trio,* *Triangle*	Difference	✓	×	×	×
Ranking	Difference	✓	✓	×	×
FPA, TPA, QDA	Descriptive Scaling	✓	✓	✓	×
Magnitude Estimation	Descriptive Ratio Scale	✓	✓	✓	✓

Table 1.1: Type of sensory testing method and the information obtained.

When selecting a sensory testing method it is necessary to consider the product's requirements and availability, subject qualifications and availability, and the logistical and cost constraints associated with the test objective.

Product Availability & Requirements

Although the amount of sample presented to subjects is often limited by the quantity of raw material available, the amount should be as constant as possible, adequate for the task, and allow for re-tasting if deemed necessary by the subject (Larmond, 1987). The Sensory Evaluation Committee of the American Society of Testing and Materials (ASTM, 1968) recommends that each panellist receive at minimum 1 oz (28gm) of a solid for use in a discrimination test, and double that amount for preference testing. Further, in order to obtain meaningful results, the samples that each subject receives must be representative of the product and be physically uniform for each treatment (Larmond, 1987).

The homogeneity and the stability of the samples or product being tested also require consideration. Unlike the visual and auditory modes, where exactly comparable stimuli can be produced, it is impossible to produce two physically identical food products such as apples or fish for all subjects to evaluate, and therefore some of the random variability attributable to subjects may be associated to stimulus variability (Land & Shepherd, 1984). The flavour, texture, and appearance of red meat, for example, with varying degrees of fat marbling and connective tissue, as well as the changes in colour that occur with the oxidation of the pigment myoglobin from deoxymyoglobin (purple) through to oxy-myoglobin (cherry red) and finally to metmyoglobin (brown), need accounting for when processing, preparing and presenting samples for evaluation.

Logistical & Cost Constraints

With all test objectives in sensory science there are logistical and cost constraints particular to each experimental method that have to be determined and met prior to execution. These may include:

1. Facilities and tools: the testing facility, its size and appropriateness; the need and availability of storage (constant high/low temperature-humidity); preparation tools and processing equipment.
2. Transportation: the transportation of products/samples from the place of production to the place of testing may be needed.

3. Human resources: staff to conduct the experiment and subjects to participate in the experiment; the need for staff and/or subject training; consultants, statisticians, butchers, chefs, survey staff, and translators may also be required.
4. The product: its size, and number. Also a product may require harvesting, capture, slaughter (possible ethical and religious considerations), manufacture, processing, baking etc.
5. Analysis: the time and cost of any physico-chemical testing; statistical analysis.

The available budget will also restrict the test objective, the testing method, the number of subjects and their level of training, the quantity of sample, and all experimental design factors such as how data is to be collected (on paper or electronically) the staff available (cutters, presenters), whether the analysis can be outsourced, and how many times the experiment can be replicated.

Subject Qualifications & Availability

People who perform sensory evaluations (also referred to as assessors or judges) can be categorised as subjects, selected subjects, panellists, or experts. Subjects are any unqualified persons involved in a test; selected subjects have been tested, trained and chosen for their proven ability; a panellist is a member of a select group with no particular expertise or abilities; and an expert is someone with considerable experience and proven ability in the assessment of a given product (Land & Shepherd, 1984). The availability of subjects, either un-qualified or qualified, will determine the type of sensory testing that can be conducted.

Threshold concentrations that elicit a taste sensation can vary by up to two orders of magnitude between subjects, and thus inter-individual variation in subject responses can result from the presence or absence of specific receptors (genetic makeup); the sensitivity to, and discrimination between, stimuli (sensory experience); redundancy, duplication and nuances associated with the terms used to describe the sensory experience (semantics); and the cognitive transformation of sensory inputs into a quasi-numeric form (reporting) (Frijters, 1984; Plattig, 1984; Brown et al., 1996). Intra-individually, responses can vary according to the time of day, hunger and satiety, hormonal influences, and age (Plattig, 1984). Limits to the magnitude of a physical difference between two stimuli that can be perceived can result in a Type II error where two physically different stimuli from two different products can elicit two identical responses, or, alternatively a Type I error where subjects respond differently to the same stimulus (Frijters, 1984).

In sensory evaluation, subjects and panels are the analytical instrument, and, as with all instruments, they require calibration to ensure they are as objective, accurate, and precise as possible. Where feasible subjects should be screened, trained, and the reproducibility of their evaluations examined (Larmond, 1987). Pre-screening, using the Munsell Colour Vision or the Ishihara tests for sensitivity and blindness for example, or threshold testing, enables those panellists with colour vision defects, or insensitivity to a characteristic or stimulus of interest to be identified and disqualified from a particular test where they would bias, skew or corrupt results. Training panellists assists to develop familiarity with the product and its characteristics, to develop a common language to describe these characteristics, to identify differences, and improve and standardise the consistency of results. Untrained subjects often struggle to disregard personal preferences and, despite understanding the terms used, do not use them in a consistent manner leading to scattered responses and statistical non-significance (Larmond, 1987). Apart from the pre-screening and training of subjects, efforts to minimize this variation and standardize the judgements of subjects include repetition, and statistical modelling (Brown, et al., 1996).

Measurement Techniques

Scales are the tools subjects use to express their perceptions, and knowing the properties and limitations of the measuring instrument is of vital importance (Land & Shepherd, 1984). The sensory scientist needs to consider the test objective, the product, and the subjects when deciding upon a response scale as inappropriate scales that are not optimised for subject use in a particular task can result in poor response data and lower motivation levels (Land & Shepherd, 1984; Stone & Sidel, 1993).

Scales used for the rating or scoring of samples are a continuum divided into equally spaced, successive values that can be graphic, descriptive or numerical. They can be uni-polar with a zero at one end or bipolar with antonyms at either end (Land & Shepherd, 1984). Too broad a scale loses discrimination but too fine a scale can introduce error.

According to Stone & Sidel (1993), in order to derive the most value from a response scale it should be:

a. *Meaningful to Subjects:* the words used must be familiar, easily understood, readily related to the product and the task, and unambiguous to the subjects. Jargon and technical terms familiar to a researcher may be meaningless to an untrained panellist.

b. *Uncomplicated to Use:* the task and scale must be easy to use. If complicated, measurement error can increase and product differences not detected.

c. *Unbiased:* it is important that the scale be balanced and that all number and word categories are equally represented so as to not influence the test outcome.

d. *Relevant:* chosen scales should only measure the intended attribute, characteristic or attitude, and not combine an element of quality and preference for example.

e. *Sensitive to Differences:* the length of the scale and the number of categories can influence the sensitivity for measuring differences.

f. *Provide for a Variety of Statistical Analyses:* to determine whether effects are a result of chance or an applied treatment the response scales need to be amenable to statistical analyses. Less flexible and less sensitive scales limit inferential power.

Stevens (1951) proposed four categories in which all response scales fall. Nominal scales for naming and classification, ordinal scales for ranking, interval scales measuring magnitudes that are equidistant between categories, and ratio scales.

Nominal scales detail only class association or recognition with no quantitative relationships existent between classes. The main feature of nominal scales is the total independence of the order between categories, and although considered a low-order scale, the assignment of ranks or percentages based on frequencies to nominal categories allows the use of statistical methods available to ordinal data. Valid analytical techniques include frequencies of occurrence, modes, chi-square, and contingency correlation (Land & Shepherd, 1984; Stone & Sidel, 1993).

Ordinal scales denote the increasing or decreasing nature of an attribute or class without providing any magnitude or distance between non-interchangeable categories (Land & Shepherd, 1984). When used in multi-product tests, all products must be examined before judgements are made and therefore sensory fatigue, which is confounded by the interactions between products which have lingering flavours or odour, can be minimised. If subjects are not trained or qualified to perform evaluations of a certain product's attributes, such as the ranking of flavour intensity, there can be no assurance that the subjects actually perceived and were able to rank flavour (Stone & Sidel, 1993). Valid analytical techniques of ordinals ranked data include all those available to the analysis of nominal data as well as the non-parametric Wilcoxon signed rank test, Mann-Whitney, Kruskal-Wallis, Friedman Two-Way ANOVA, and Kendall's coefficient of concordance (Land & Shepherd, 1984; Stone & Sidel, 1993).

An interval scale is defined by equi-distances between categories on a scale which has an arbitrary zero point, and, therefore, is unable to measure the absolute magnitude of an attribute (Stone & Sidel, 1993). Mathematical comparison is possible using the arithmetic mean, standard deviation, and a variety of non-parametric and parametric analyses such as Pearson correlation factor analysis, or discriminant analysis (Land & Shepherd, 1984; McDowell, 2006).

Ratio scales differ from interval scales in that a constant ratio exists between points on a scale and zeros are absolute therefore possessing geometric properties rather than just the arithmetic properties (Land & Shepherd, 1984). Similar to natural ratio scales, there are no arbitrarily limited endpoints, and, as the choice of numbers will vary amongst subjects, the scores are transformed so that the geometric mean of all subject data equals 1.00 and the logarithms are analysed by analysis of variance (Larmond, 1987). In moving from nominal to ratio scales the demands on subjects increase both in their comprehension of instructions and their ability to respond.

In order to minimize potential variations in the interpretation of scale values by subjects, sensory scientists sometimes develop and use standards which are then used to train subjects (Land & Shepherd, 1984). As described by Cardello and Maller (1987) the question of the validity of the scales of sensation is a ‘thorny’ one, where no one scale can be considered better than another when differing methods of judgement are used.

Experimental Design

Experimental design is an organized approach to the collection of data and requires population definition, randomization, administration of treatments, consideration of the sample size requirement, and sound statistical analysis (Gacula, 1988). The chosen design will affect the accuracy of results and the feedback into the test objective. A well designed experiment reduces costs, simplifies the interpretations of the results, and yields useful and meaningful outcomes (Larmond, 1987; Gacula, 1988).

The number of subjects used in a sensory test will influence the statistical significance of the results with too few a number used requiring subject responses to be considerably different between treatments in order to produce a significant result (Land & Shepherd, 1984; Larmond, 1987). Although there is no set number, consumer preference/acceptance testing requires the number of subjects often to be in the hundreds in order to cover demographic, regional, and cultural differences. When using trained or expert panellists Larmond

(1987) states a panel numbering 10 subjects is common and that a 5 subject panel is the recommended minimum number.

According to Gacula, (1988) when accounting for Type I and Type II errors and desired test sensitivity, it is possible to determine appropriate sample sizes using the following formula:

$$n = [(Z_\alpha + Z_\beta)^2\sigma^2]/(\mu_1 - \mu_2)$$

where Z is the area under the curve of the standard normal distribution; σ^2 is the variance and $\mu_1 - \mu_2$ is the desired difference to be detected. Gacula (1988) noted that in order to detect a difference of 0.5 on a nine point scale (assuming σ= 1.0) with $\alpha = 0.05$ and a power of 0.90, 52 panellists would be required per treatment. By changing the detectable difference value from 0.5 to 0.4 and 0.6 increases and decreases the required number of panellists to 81 and 36 respectively.

It is considered ‘best practice’ to run two sessions in one day with each subject participating in both sessions to obtain replication in the judgements, provide a measure of consistency of the panellists, and to minimise exposure to systematic errors (Larmond, 1987).

One important aspect of sensory testing is the number of samples that are presented to subjects for evaluation. Physiological fatigue of the sensory organs and/or tiredness, boredom and confusion of the subjects can result from presenting too many samples for evaluation (Land & Shepherd, 1984). Determining the appropriate number of samples relates to the type of stimulus being evaluated and the complexity of the task. For example, the visual evaluation of samples is less taxing on subjects when compared to the evaluation of a sample's flavour or texture. The swallowing of samples rather than being able to spit them out and the need to taste unpleasant samples can affect both the required evaluation time and subject enthusiasm (Land & Shepherd, 1984). The expectations of subjects regarding the number of evaluations, the number of samples, the reason for doing the test, the importance of the test and their relationship to the experimenter can also affect subject performance (Land & Shepherd, 1984).

Environment, Sample Allocation, & Presentation

To assist the subject to concentrate at the task at hand and provide the optimum setting for unbiased judgement, subjects should be alone in an environment with negligible distractions and interruptions (Land & Shepherd, 1984; Larmond, 1987). The testing area should be quiet, comfortable, free from foreign odour (maintaining positive pressure in the testing room), and preferably be maintained at a constant temperature and humidity. Individual booth's with neutral colours and uniform lighting that does not distort the colour of samples is considered the optimum environment (Larmond, 1987).

Samples need to be prepared in an area with sufficient counter space to allow for their efficient allocation and serving. Preliminary testing is usually necessary to determine efficient methods of preparation and allocation

(Larmond, 1987). Samples should also be served at the temperature in which they are normally consumed and held constantly at that temperature prior to and during testing. If held for a long period of time precautions should be taken to prevent samples from drying out and any quality changes from occurring (Larmond, 1987).

All samples should be individually coded as subjects may learn, or believe they know the identity of a particular sample and start to rate it consistently according to its label and not its attributes (Land & Shepherd, 1984). The code assigned to samples should not introduce any bias or provide any indication of the identity of the treatments, and, according to Larmond (1987) three digit random codes from random number tables are the most appropriate and are widely used.

Serving containers that are identical and that do not impart any taste or odour to the product should be chosen for sample presentation. Colourless or white containers are best in order to not mask any colour differences between samples (Larmond, 1987).

Psychological Error and Allocation (Randomization)

As the subjects of a sensory test are human they are prone to a number of well documented psychological errors when conducting sensory evaluations. Even a well trained panel can respond poorly if the experiment is designed poorly and psychological errors are not taken into account. These errors include: time-order errors; errors of central tendency; errors of expectation, habituation and anticipation; stimulus errors: logical and leniency errors; halo effects; proximity errors; contrast and convergence errors; and range-frequency effects.

The first-sample effect is a time order error where the first sample or product presented in a multi-sample or multi-product test is often evaluated higher on measurement scales than when the same sample or product is evaluated later in the series of presentation (Stone & Sidel, 1993). To control for this phenomenon each sample of a multi-sample test should come first an equal number of times in the sampling series.

Errors of central tendency occur when subjects avoid both poles of a bi-polar, and the upper end of uni-polar response scales, and tend to score around the central position. This can result in products or samples being less differentiated than otherwise would occur with trained subjects or subjects familiar with the range of stimuli being evaluated (Land & Shepherd, 1984; Stone & Sidel, 1993).

Errors of expectation, habituation and anticipation occur when prior knowledge or experience generate expectations within a subject for specific attributes or differences between samples and products. Habituation occurs when multiple stimuli nullify the subject's perception of an actual change in sample or product. Anticipation errors occur if there is a perceived change when in fact no actual change in sample has been presented to the subject (Stone & Sidel, 1993).

Stimulus errors occur when subjects respond in an atypical manner owing to the fact that they have or believe that they have some prior knowledge on the products or samples being used in the test, and, according to Stone & Sidel (1993), reinforces the notion that participants involved in the setting up of the test should not then become subjects.

Logical errors arise when subjects follow their own logic in determining the requirements of the task at hand and results from unacquaintance with the protocols of the employed technique. Leniency errors are the result of the subject allowing their feelings toward the experimenter influence their scoring of attributes or when subjects try to comply with what they feel the experimenter desires (Land & Shepherd, 1984; Stone & Sidel, 1993). 'Double blind procedures' attempt to minimise these effects by using an experimenter unfamiliar with the trial objective, the classification of subjects, or the treatment conditions (Land & Shepherd, 1984).

The halo effect results when a response to one particular question by a subject influences successive responses by that subject and are common with untrained panellists or consumers who are attempting to justify a preference rating or are suffering from physiological fatigue due to numerous re-tastings (Stone & Sidel, 1993).

As the name suggests proximity errors are said to occur when attributes measured in close proximity to one another are scored more similarly than those attributes that are measured farther apart during evaluation. Possible solutions include randomizing the order in which the attributes are assessed by subjects.

Contrast errors can occur, for example, when a product of high intensity in a particular attribute is immediately followed by a product of low intensity. This can result in the scored difference being far greater than the actual difference, with the low intensity product score being more exaggerated than if the preceding product was closer in intensity. Convergence is the opposite effect where like products are presented in close proximity (Stone & Sidel, 1993).

Both the range and frequency of presented stimuli can influence the evaluations made by subjects. Similar to the error of central tendency subjects adjust the centre of the rating scale in the direction of the centre of the stimulus range (Land & Shepherd, 1984).

Proper experimental design is used to decrease the risk of introducing bias, and randomization is one of the fundamental principles of good experimentation (Piggott *et al.* 1998). To account for psychological errors the order of presentation of samples to subjects should be randomized or balanced with the ideal situation resulting when every possible order occurs an equal number of times (Larmond, 1987). It is also necessary to randomly allocate experimental materials to treatments so that each has an equal chance of being assigned to a particular treatment in order to guarantee that a statistical test will have a valid significance level (Gacula, 1988).

Data Analysis & Statistical Considerations

The physical and sensory complexity of some foods can result in as many as forty sensory attributes being measured, as well as numerous instrumental measures (Cardello & Maller, 1987). The power of multivariate statistics to deal with such large numbers of variables is often employed to expose the underlying physical and perceptual dimensions (Cardello & Maller, 1987).

Non-parametric tests, according to Stone & Sidel (1993) provide the sensory professional with additional tools for data analysis when there are reasons to justify their use, otherwise, in agreement with O'Mahony (1986), these authors note that parametric methods are preferable owing to the fact they use scale data obtained from the subjects.

As with all data sets the reliability, validity, and the amount of replication used to gather data underpins analytical results. In sensory science, reliability refers to the ability of subjects to respond repeatedly in the same manner from the same stimulus. Low subject numbers make it important that the experimenter has confidence in the ability of those subjects to reliably respond. Validity refers to the accuracy of those responses. It is of no value to the sensory professional if subjects are repeatedly responding to stimuli in an invalid manner. Although difficult to assess, two types of validity are commonly referred to in sensory science – face and external validity. Face validity is said to prevail where responses are in line with expectations, and external validity occurs when a different and/or larger set of subjects respond in accordance with the original set of subjects (Stone & Sidel, 1993).

Independent re-evaluation in identical experimental conditions that allows for the consistency of individual subjects and panels to be determined is the goal of replication (Piggott *et al.* 1998). Practical considerations such as product availability, preparation requirements, product stability, subject availability (both in numbers and frequency), and the information required all affect the ability to replicate (Stone & Sidel, 1993). Replication in foods such as meat is not a simple task as the individual animals themselves, the place and method of slaughter, and the storage time and temperature variables introduce variability (Piggott *et al.* 1998). Products, according to Stone & Sidel (1993), are often more a source of variability than subjects, and can be out of the control of the sensory professional.

Concordance Analysis, an application of Principal Components Analysis, is a powerful tool for analysing the performance of sensory panellists, and can reveal whether panellists agree or not, panellists that cannot reproduce their evaluations, and distinguish panellists who have problems with a particular attribute.

Sensory-instrumental analysis compares and contrasts a data set containing a collection of sensory assessments on a number of products with a data set containing a number of instrumental measures on the same products (Dijksterhuis, 1997). According to Dijksterhuis (1997), multivariate methods to study the relationships between sensory and instrumental data sets can differ in three respects – symmetry, measurement level, and criterion.

The symmetry refers to the way in which the data is treated by the analytical method. Asymmetric methods, that include Partial Least Squares Regression, Principal Components Regression, Redundancy Analysis and Multiple Regression, attempt to predict one data set from another and treat both data sets differently (Dijksterhuis, 1997). Symmetrical methods, including Canonical Correlation Analysis and Procrustes Analysis, investigate only the relationships between the data sets with neither set used as the object of prediction (Dijksterhuis, 1997).

Non-linearities at the measurement level in sensory-instrumental analysis presents the experimenter with the difficult task of finding the right balance between imposing linear restrictions with the risk of missing interesting relations, and imposing hardly any restrictions with the risk of fitting noise (Dijksterhuis, 1997). The criteria in which the relationship between the sensory and instrumental data sets are defined differ according to the multivariate procedure, and can be based upon maximal covariances, maximal correlation or minimal variances (Dijksterhuis, 1997).

Principal Components Analysis (PCA) is an often used statistical technique in sensory science and functions to reduce a set of individual items (variables and data) into components. The first component has maximum correlation with all variables in the data set and accounts for the greatest amount of variance, the second component accounts for the second-largest amount of variance etc. This trend continues until all variation as practicable has been accounted for

(Powers, 1984). Finally, asymmetric Partial Least Squares Regression analysis (PSLR) is often used to examine the relationships between data sets, such as those between instrumental (x) and sensory (y), by predicting one from the other, as well as attempting to find the ‘best’ solution of *X* that will explain variations of the *Y* variable set (Chung et al., 2003).

An understanding of the principles of sensory evaluation and the techniques available for the analysis of food makes it possible to consider and develop a sensory testing protocol for a raw fish product such as Southern bluefin tuna. Once such a tool is at hand it would be possible to measure the affects of variations in, or manipulations of, any on-farm or in-chain practices on the sensory profile of the end product. Furthermore, if linked to consumer preference or acceptance analyses, the results could help to identify practices that maximise the product characteristics associated with a ‘good quality eating event’ – the reason d’être of all farm managers.

1.4 Aims & Objectives

The first objective of the Japan-based research effort was to identify, investigate, and compare the instrumental techniques used by researchers at the Tokyo University of Fisheries to measure the qualities of fish flesh with those used by the tuna flesh quality research team in Australia. Secondly, in collaboration with Japanese researchers and industry representatives, to develop new, and appropriate, instrumental and sensory techniques for the measurement of the qualities of Southern bluefin tuna – reviewed in the methods section.

The flesh qualities of the product in Japan, the effects of on-farm and in-chain manipulations on selected quality characteristics of the end product could all be examined using sensory and instrumental techniques. Chapter three examines the effects of using vitamin supplements as per the industry practice on the flesh concentrations of vitamins and the colour shelf-life of the end product. Chapter four examines the effects of an industry harvesting practice on the sensory and biochemical characteristics of quality. Finally, chapter five investigates the time-temperature management of air-freighted SBT to Japan, as well as the subjective and objective quality associated outcomes of two industry shipping procedures.