INTRODUCTION

CHAPTER ONE

This thesis describes the development and validation of a competency based assessment tool to evaluate speech pathology students' performance in the workplace. The research addresses a need that was identified by both the university programs involved in educating speech pathology students and the professional association, Speech Pathology Australia.

The original proposal successfully submitted for Australian Research Council funding support by the University of Sydney and Speech Pathology Association of Australia (SPAA) in association with The University of Newcastle and Charles Sturt University. This proposal identified the need for a valid and reliable national assessment tool to measure the clinical performance of speech pathology students. A competency based approach was expected given that the profession had already developed a competency based description of the work undertaken by speech pathologists in Australia known as the Competency Based Occupational Standards (Entry Level) or CBOS (SPAA, 2001)[1].

The CBOS is the foundation for the SPAA accreditation processes that evaluate Australian speech pathology programs and subsequently deem students graduating from these programs to be eligible for membership of the association. Each university program has developed their own workplace assessment formats, either directly or indirectly based on the CBOS, but none of these tools has been evaluated for their reliability and validity. This burdened clinical educators (CEs) in the field with the need to be familiar with the use of multiple assessment tools and possibly impacted upon the quality of the assessments and subsequent outcomes. For example, field CEs providing placements for students in New South Wales were accommodating students from 4 different university programs and similar situations were likely to develop in other states as well. In addition, the variety of unvalidated assessments did not provide SPAA with assessment information in which they could have complete confidence when determining graduating students' eligibility for membership.

The funding support provided by the Australian Research Council in collaboration with Speech Pathology Association of Australia enabled the process of developing a validated assessment tool of speech pathology students' performance in the workplace to be undertaken. It soon became clear that this was a unique endeavour in speech pathology and indeed allied health and medicine, with the exception of the Clinical Performance Instrument developed and validated by the American Physical Therapist Association (Roach et al., 2002).

[1] The term 'CBOS' throughout this thesis refers to the documented titled 'Competency Based Occupational Standards(CBOS) for Speech Pathologists (Entry Level)', revised and published in 2001 (SPAA, 2001).

Ominously titled articles on the topic of assessing workplace performance were found such as: “The long and tortured history of clinical evaluation” (Woolley, 1977), suggesting that perhaps the lack of validated workplace performance assessments of health professionals was due to difficulties inherent to the task. Indeed, as stated by Kane (1992):

> “So, competence assessment looks easy. Difficult tasks that look easy tend to be frustrating.” (pp. 164)

To better understand the nature of the task and the unique challenges and constraints involved, an extensive literature review was undertaken regarding the nature of competency and its relationship to professional practice. The purpose and nature of assessment and validation of assessments was also examined, both in general and specifically in relation to workplace performance. This information is described in the first section of the thesis, Chapters Two and Three.

The understanding gained from the literature review led to the development of the research over two phases. The first phase involved development of the assessment tool format and processes with careful attention being paid to integrating multiple sources of information regarding speech pathology competence and assessment of workplace performance. This included knowledge derived from research, theory, expert opinion, and the opinions of those most intimately involved in the effort to ensure students graduate sufficiently competent to practise: the students themselves and the CEs who work so closely with them. Phase 1 (Chapters Four and Five) describes this developmental process.

The second phase involved field testing the instrument developed as a result of Phase 1 and the methodology involved is described in Chapter Six. Once field testing was completed the evidence required to evaluate the research assessment tool’s validity was identified and analysed and is documented in Chapters Seven and Eight. Chapter Nine discusses, evaluates and weighs up the evidence derived from this process and offers an opinion regarding the tool’s validity, its strengths and weaknesses, and possible lines of future inquiry. Most importantly, the potential usefulness of the assessment tool for the purpose for which it was designed is described.

ISSUES IN COMPETENCY BASED ASSESSMENT

CHAPTER TWO.

2. THE NATURE OF COMPETENCY

2.1. Why Assess Competency?

At first glance it would appear that the reasons for assessing professional competency are self-evident and relatively straightforward. Simply: it is important to know if a speech pathology graduate is competent to practice as a speech pathologist. However, why it is imperative to assess professional competency depends on one's viewpoint and may affect how competence is defined and therefore assessed. The public, the faculty, the profession, and students all have reasons why they believe competency should be assessed and Bargagliotti, Luttrell, & Lenburg (1999) propose that each feels they have the most invested in appropriate assessment. It is accepted that there are three main functions or reasons for assessing competency: fitness for *practice*, for *award* and for *purpose* (Cross, Hicks, & Barwell, 2001; Education, 2001; Priest & Roberts, 1998; QAAHE, 2001).

Assessment for fitness for *practice* aims to ensure acceptable standards for effective practice of the profession are maintained. This viewpoint is generally held by the profession and reflects the profession's concern regarding maintaining quality of services and protecting those who seek services from the profession. Eraut (1994) also argues that every profession has a need to maintain their status and reputation by exclusion and the definition of competency and its assessment is driven by this need. For example, an important impetus for the speech pathology professional association in Australia to define their understanding of competency has been the need to assess competency to ensure overseas qualified speech pathologists meet the standards of practice held by the speech pathology profession in Australia (Dawson, 1993).

Assessment for fitness for the conferring of a particular *award* endeavours to determine whether the students should receive the qualification for which they are studying. This is of particular concern for universities who function in a competitive tertiary education sector and need to ensure they have a reputation for graduates of high quality. In addition, universities assess to discriminate between different levels of achievement and to inform the learning and teaching process (Cross et al., 2001).

Assessment for fitness for *purpose* ensures that graduates possess relevant knowledge and skills and therefore ability to practice. The employer needs to know that the graduate is able to do the job. The UK Academic and Practitioner Standards (QAAHE, 2001) also suggest that this should not only include the graduate's fitness for their first post but also for continuing their professional development. Eraut (1994) also argues that the government (who is also frequently the employer of speech pathologists in Australia) is concerned with assessing fitness for purpose with a view to limiting professional autonomy to safeguard the interests of the public. This has certainly been the case in Australia where the definition of speech pathology competencies was given impetus and support by Commonwealth Government initiatives to promote a competency based approach to ensure various occupations and professions were able to meet the needs of their customers and clients (Dawson, 1993).

In addition, Australian university speech pathology programs are accredited by Speech Pathology Australia (SPAA), the national professional association, so that their graduates are eligible for membership to the professional association. To be accredited the programs are required to demonstrate that their graduates have been adequately assessed against all the units and elements of competency represented in the Competency Based Occupational Standards – Entry Level or CBOS (SPAA, 2001). The graduates of an accredited program are then eligible to apply for membership of SPAA. This accreditation process and the publicly available CBOS document (SPAA, 2001) are a de facto protection of the interests of the public who are not safeguarded by legislated registration of the profession of speech pathology, except in the state of Queensland.

While assessing for fitness to practice, for award and for purpose are viewed from slightly different positions – the public, the employer, the profession, the educator, or the student themselves – they are clearly closely aligned. These complementary but multiple purposes are encapsulated in the stated need to develop and implement defensible assessments of competency that demonstrate that graduates can perform 'on the job' (Carraccio, Wolfsthal, Englander, Ferentz, & Martin, 2002; Swchwabbauer, 2000; Whitcombe, 2002).

Thus it is the case that competency should be assessed simply because it is important to know whether graduate speech pathologists are equipped to competently practice as speech pathologists. However, there are many viewpoints as to why competency is important and why it should be assessed, and the beliefs of the various stakeholders will have an impact on how entry level competency is defined.

2.2. Competency

2.2.1. Defining Competency

2.2.1.1. Describing Competency

Defining competency is the first step in developing an assessment of competency. However, the term “competency” is so commonly used that it is difficult to define (Eraut, 1994). The Macquarie Dictionary (Delbridge et al., 1981) defines “Competent” as:

> Competent *adj.* 1. properly qualified, capable. 2. fitting, suitable, or sufficient for the purpose; adequate. 3. rightfully belonging; permissible (fol. by *to*). 4. *Law.* (of a witness, a party to a contract, etc.) having legal capacity or qualification.[L *competens*, ppr., being fit] (pp. 387)

Eraut (1994) suggests that it depends on the circumstances as to whether describing someone as competent is positive i.e. they can get the job done or negative i.e. adequate but less than excellent. This issue is highlighted by commentaries on competency and university educated professions such as that of Wilson (1992) when Wilson states that universities see their goal as being excellence, not ‘just’ competency. Eraut (1994) further proposes that professional competence has at least two dimensions. First, ‘scope’ relating to the range of roles, tasks and situations in which competence is expected. Second, the dimension of ‘quality’, in that professional competence should be seen as existing on a continuum from novice to expert rather than a binary scale (competent or not).

The scope of tasks for which a graduate is expected to be competent and the quality of performance expected will be different from one profession to another (Eraut, 1994). This will be determined by the shared assumptions and traditions the profession and workplace hold regarding what can be expected from a new graduate and will determine the way work is organised and allocated when they enter the workplace. In the case of speech pathology, graduates are expected on graduation to be relatively autonomous and to be competent such that they can handle most situations across the whole scope of practice. However this expectation of competence is accompanied by the caveat that the graduate should have access to professional supervision or mentoring and in-line managerial supervision. This support will

need to be accessed when working in situations where the combined features of the client or speech pathology service create complexity (SPAA, 2001).

The speech pathology profession in Australia has also outlined the range of roles, tasks, and situations in which all speech pathologists are expected to be competent. The process of determining this scope will be discussed elsewhere, as well as determining the level of competence at which a speech pathologist would be considered competent. The issues for discussion at this point are: how does one define competence and what are its components?

Defining competency is complicated by the fact that competency is intangible, it cannot be directly observed only inferred from demonstrated behaviours (Gonczi, 1992). Gonczi suggests that the construct of competence is an attempt to describe the multitude of personal characteristics or attributes that underpin and enable an individual to perform competently in their occupation. Further, some personal attributes that underlie professional competence are easily recognisable but others may be poorly defined, understood or not even recognised (Gonczi, 1992).

The literature on competency generally defines competence by the behaviours from which competence can be inferred and recognised. Generally two types of observable competence are represented in the literature, either separately or in combination. First, competence is seen as a purely technical matter i.e. can the individual perform the specific job tasks or not? This approach focuses on specific, measurable descriptions of the job tasks. Second, competence is conceptualised as a set of underlying traits that enable a person to perform competently and are inferred to be present as demonstrated by the person's competent performance. This approach leads to an attempt to describe these underlying traits, separate from the occupational tasks themselves. The third option is to see competence as a combination of both. Conceptualising or defining competence is a critical first step towards assessing it therefore these three theoretical viewpoints will be explored.

2.2.1.2. Competence as Technical Competence

Conceptualising competence as being able to perform a specific set of tasks to a certain level of skill is rooted in the behaviourist tradition (Eraut, 1994), which has generally focussed more on training than qualifications, and as such has often been applied to technical or skill based occupations such as trades. However, breaking down an occupation into specific component skills and assessing each separately ignores the general complexity of any workplace environment (Ling, 1999). Thus defining competence in this way is generally

criticised as being too narrow (Eraut, 1994; Grant, 1999; Harris, Guthrie, Hobart, & Lundberg, 1995; Wolf, 1995). However, it is attractive because it is easier to conceptualise and observe than assumed underlying competencies (Harris et al., 1995).

Such task based approaches generally result in exhaustive, atomised descriptions of behaviours or skills that can be observed. This was the experience of regulatory authorities of the United States in the 1970's who introduced certification of teachers through competency assessment (Wolf, 1995) and paradoxically caused a lowering of the standard of teaching. Wolf (1995) also argues that the National Vocational Qualifications system in the United Kingdom attempts to define competency to unattainable levels of precision and therefore is unwieldy and neglects important, abstract aspects of competence.

Concentration on descriptions of the skills in an occupation neglects the holistic nature of work and those competencies that must be inferred from observable action, such as knowledge and attitudes. These competencies can be equally argued to be essential to competent job performance (Eraut, 1994; Harris et al., 1995; Wolf, 1995). Chapman (1998) epitomises this dilemma when she states:

> "Professional practice is comprised of both observable actions or behaviours and professional attributes that are implicitly related to that performance. Do we then emphasise the assessment of technical skills because these can be easily defined and marked, or should we explore the abstract issues more closely in an attempt to make them explicit for the benefit of the educator who is marking as well as the student? Will this analysis of the elusive concepts reduce it to parts, which, when measured and aggregated again, would not reconstitute the whole we were attempting to assess?" (pp. 163)

In addition, there are concerns that assessing for the skills required for the job today does not reflect the dynamic nature of the workplace and the need for employees to transfer and apply skills and knowledge to new situations and environments (Harris et al., 1995). This is seen as being particularly pertinent for complex work environments (Down, Martin, Hager, & Bricknell, 1999).

2.2.1.3. Competence as Generic Competence

Conceptualising competence as a set of underlying traits or generic competencies is attractive. Eraut (1994) states that the generic approach to competence has developed from

management research and is focussed on what enables a person to do their job, including personal qualities. These are generally seen as being predictive of the person's ability to continue to perform competently and may include qualities such as initiative, persistence, and critical thinking (Eraut, 1994). However it is self-evident that there is a skills based component to competence in any occupation, including the professions, and that these need to have been developed to a competent level prior to commencing employment.

A related way of conceptualising competence, based in cognitive psychology and linguistic constructs, is to see competence as being different from performance (Eraut, 1994). Competence is seen as something the person knows and can do in ideal circumstances, whereas performance is what they actually do in the situation under observation. Thus performance becomes something that includes accessing and using underlying knowledge and abilities and other affective, motivational, attentional, and stylistic factors that will influence the final response.

This concept appears to underlie assessment strategies promoted by Bargaliotti et al. (1999) and Luttrell, Lenburg, Scherubel, Jacob, & Kock (1999). Their assessment of practical competence in nursing is based on carefully structured Competency Performance Evaluations (CPE) conducted in the most stable and controlled environment possible with the stated intention to ensure that the assessment is fair. Otherwise, they claim, it would be 'akin to administering a final course examination in the midst of the student cafeteria or during a fire drill' (Bargagliotti et al., 1999). Presumably the intention is to enable students to perform as closely as possible to their underlying competence.

Competence as an underlying quality demonstrated to a greater or lesser degree by actual performance is also inherent in Miller's model which is frequently cited in medical education (Miller, 1990). This model proposes a pyramid with "knows" or knowledge at the base, "know how" or competence on the next layer, "shows how" or performance next, with "does" or action at the apex. Thus you can have sufficient knowledge, judgement or skill to be competent, and your performance is linked to this, but performance is seen as something that you do when you are faced with a client (Miller, 1990).

One of the primary difficulties of conceptualising competence in this way is that the relationship between the defined 'underlying competence' and actual competent performance is not clear. If one uses the strategy suggested by Bargagliotti et al. (1999) and Luttrell et al (1999) and described above, it can be argued that the students may not be able to function competently in the complexities of the 'real life' ward environment and therefore does not in

fact assess competence. This situation reflects the lack of clarity regarding what 'competence' actually is.

More recently the field of medical education and assessment have also acknowledged that the relationship between competence and performance, as conceptualised by Miller (1990), is complicated and that competence is a necessary but not sufficient requirement for competent performance (Rethans et al., 2002; Schuwirth et al., 2002). Schuwirth et al. (2002) define competence as how people perform in ideal conditions knowing that they are being challenged to demonstrate that they have the knowledge, skills, and attitudes required for a task. On the other hand, performance describes how people behave when in real life situations and when they are not being observed. This performance will be influenced by everyday constraints including internal factors such as the professional's own health and external factors related to the context they are working in e.g. workplace constraints. It is suggested the emphasis of assessment should be on assessing performance rather than competence (Rethans et al., 2002; Schuwirth et al., 2002). This is very much in alignment with the three main reasons for assessing competency: fitness for practice, for award, and for purpose.

At first glance this seems to simplify the task of conceptualising competency in that deciding that someone is competent to practice is simply a matter of observing his or her real life performance. However, the question still remains as to what aspects of their performance should be observed and what does competent performance look like? In the context of this research it is this 'competent performance' that is the focus, not solely assessment of competence as an underlying quality in the absence of real life, work based performance.

2.2.1.4. Competence and Competent Performance

The majority of the literature on professional competencies and their assessment conceptualises competent performance as a combination of performing specific occupational skills to a desired level of performance and possessing generic or underlying competencies (as inferred from behaviour). The relative weighting given varies but it is clear that both are valued. This combined approach was initiated, and its application to the professions consistently promoted in Australia, by bodies such as National Office of Overseas Skills Recognition (NOOSR) and the National Training Board (NTB) from the early 1990's.

The Australian approach places significant weighting on underlying competencies. For example Gonczi (1992) suggests that competence should be inferred through a combination of attribute and performance based inferences. Attribute based inferences are founded on the

definition of personal attributes believed to underlie competent performance e.g. skills, knowledge, and attitudes. If these attributes are present in the person it is assumed that this comprises evidence that the person is likely to be competent. For example, in professional contexts, knowledge is assessed as a key component of competent performance.

Nevertheless, it is simplistic to assume that personal attributes will automatically translate into competent performance (Gonczi, 1992). Performance based inferences are also required and are derived from observed performance in the actual workplace. This involves an inference that competent performance reflects underlying competence e.g. underlying competence in problem solving should be evident in workplace performance. Performance based assessment requires specification of what people have to be able to do, what level of performance is required, and the circumstances in which that level of performance needs to be demonstrated.

Similarly, Eraut (1998) distinguishes between notions of 'competence' and 'capability'. Competence is inseparable from the particular workplace and describes the ability to perform tasks and roles required to a particular standard. Capability is a more individually situated quality, similar to the 'attribute' approach endorsed by Gonczi, and includes what a person can think or do if given the opportunity. Eraut makes the point that for any professional starting a new job, there will always be a part of the job competence that is outside their capability and needs to be learned, and that the employee will also bring capabilities that the job doesn't require. It's the individually situated capabilities that enables professionals to extend their current range of competence (Eraut, 1998). Finally, Eraut suggests that it is a part of a professional's capability to be able to develop or transfer one's practice as well as to create new knowledge and learn from others (Eraut, 1998).

Gonczi (1992) states that the benefit of combining both competence in work based skills and personal attributes or capabilities, is that it focuses attention on the personal attributes of competent professionals as well as how these are applied in competent performance in the actual workplace. The further advantage of including the performance aspect of competence is that it promotes the concept of the 'worker' being able to flexibly apply their competencies to new contexts, ensuring that competencies for today equip the employee to perform competently in the future. This approach was endorsed in Australia by the NTB and resulted in two types of competencies being defined (Stern, Baily, & Merrit, 1996): key competencies and functional competencies

Seven generic “Key Competencies” were endorsed by the Mayer Committee (Harris et al., 1995) representing underlying attributes that contribute to competent performance. These competencies were as follows (Harris et al., 1995; Sharpley, 1997)

1. Collecting, analysing and organising information.

2. Communicating ideas and information.

3. Planning and organising activities.

4. Working with others and in teams.

5. Using mathematical ideas and techniques.

6. Solving problems.

7. Using technology.

The second set of competencies is termed by Stern et al. (1996) as ‘functional competencies’, encompassing occupationally specific skills and expected to include some aspects of the key competencies but not necessarily all of them. Occupational specific competencies are generally conceptualised as including the knowledge and attitudes required for competent performance of an occupation (Harris et al., 1995).

Thus the two types of competencies are closely intertwined in practice. Ultimately they are combined to develop a specification of the knowledge, skills, and attitudes that underpin successful performance and generic capacities such as the ‘Key Competencies’ (Hager, Athanasou, & Gonczi, 1994). When the major components of activities in the workplace are analysed, it is generally found that they require a combination of specific work skills (occupational competencies), generic competencies (or Key Competencies), and consideration of particular aspects of the work context that require integrated performance of these competencies (Down et al., 1999).

The emphasis here is on identifying and acknowledging that successful work performance is more than performing discrete tasks successfully but also involves effective holistic integration and coordination of these tasks (Hager et al., 1994). Ling (1999) states that competency is “…composed of complex capacities to perform and to manage multiple tasks,

to deal with contingencies and novel situations, and to apply existing capacities to new contexts…" (pp. 3).

2.2.1.5. Defining Competence: Summary

In summary, a case has been made for defining competence as a combination of generic (e.g. Key Competencies) and occupational competencies (or functional competencies)[2] necessary for competent integrated and coordinated performance across the scope of a practice of a particular occupation. Thus the scope of practice, which involves identifying the roles, tasks, and situations in which competence is expected, will need to be identified. Furthermore, the level of performance required for someone to be considered competent across the scope of his or her work practice will need to be determined.

The following sections will address these issues in more detail, as well as identify how notions of competency have been interpreted and applied differently across the vocational education/training sector (VET) and universities. These differences have created debate about whether competency based education and assessment approaches are appropriate for university 'educated' professionals as opposed to VET 'trained' workers. This issue warrants consideration when attempting to develop a competency based assessment of professional performance, such as the one being undertaken in this research.

2.2.2. Competencies and University Educated Vocations

2.2.2.1. History of Competency Based Education in Australia

There is sufficient debate in the literature to indicate that the university sector has been reluctant to adopt the philosophy of competency based education (CBE) and assessment, unlike the VET sector. Admittedly the VET sector was required to do so through legislation; however it also would appear that it was congruent with its understanding of its role of training/educating people for practical, vocational employment.

[2] The term 'generic competency' will be used in the thesis to describe those competencies that are particular combinations of knowledge, skills, and personal qualities that facilitate the appropriate application and development of occupational competencies both now and into the future. Occupational competencies are those competencies that describe the nature of the profession's work and will include particular knowledges, skills and personal qualities in combination with some aspects of generic competencies to ensure competent performance. These concepts will be elaborated upon further in Sections 2.2.2., 2.2.3. and 2.2.4.

The university sector, however, has appeared to perceive the competency based approach as a 'threat' to the intellectual standards and attitudes perceived as unique to a university education (Geffen, 1992) or traditions of scholarship, intellectual inquiry, and excellence (Wilson, 1992). This is despite the fact that universities have a long history of vocational education in the areas of health, engineering, architecture, and music, which has included practical, vocationally oriented assessments (Jones, 2000).

It has been suggested that the CBE approach is politically motivated (Grant, 1999; Wilson, 1998) and constitutes an unwanted external control of professional activities. Wilson (1998) argues that CBE is solely a political move to impose economic ideologies on professional work practices such that the only knowledge and skills deemed worth developing are those related to increasing economic productivity. There is no doubt that the CBE movement has been strongly supported by government agendas in Australia since 1980; however it can be argued that this support is a representation of the employer and consumer concerns regarding competency in the workplace. Certainly Carracio et al. (2002) suggest that the CBE movement is akin to the Flexnerian revolution in medical education of the early 1900s where public concern regarding the poor education of doctors resulted in drastic reform.

CBE has been embraced in the VET sector in Australia in response to the Commonwealth Government's national training reform agenda in the late 1980's to early 1990's (Harris et al., 1995). This required defining of competencies for particular occupations and shaping education and assessment according to these competencies. The NTB (later subsumed under the Australian National Training Authority [ANTA]) was established in the early 1990's to ratify these competency standards. This approach aimed to ensure that education not only focussed on knowledge, as it traditionally had, but also on employment related generic competencies (Sharpley, 1997).

NOOSR was also established to recognise overseas qualifications on a competency basis and this necessarily embraced university educated professions. Speech pathology in Australia became involved in developing competency based standards through NOOSR funding to assist in recognising the qualifications and competency of overseas graduates (Dawson, 1993). This process resulted in a statement of competencies required for effective practice of the speech pathology profession subseqently revised and published as Competency Based Occupational Standards for Speech Pathologists – Entry Level (SPAA, 2001). It was then a small step for speech pathology, and no doubt other university educated professions, to use

competency based standards as a guideline for determining the eligibility of Australian graduates to enter the profession.

2.2.2.2. University Concerns Regarding Competency Based Education

There are a number of reasons why universities have been concerned about adopting competency based education approaches. In Australia the VET sector has traditionally been identified as post-secondary, non-university education and training, focusing on apprenticeships rather than a general, intellectual education (ANTA, 2003). CBE is seen as being congruent with vocational education but not appropriate for university educated professions (Jones, 2000) as these are thought to require specialised intellectual capacities rather than sets of practical skills (Eraut, 1994). Universities aim to develop these capacities through extending and passing on a body of knowledge in the context of an intellectual culture, perpetuating the higher education system's beliefs, values, and procedures, both to advance individuals socially or economically and to equip them for lifelong learning (Miller, Imrie, & Cox, 1998; Wilson, 1992). However, Stern et al. (1996) find that there is now a trend toward convergence of vocational and academic education as non university educated employees are being required to possess high level thinking skills and abstract theory related to their work practice as well as having lifelong learning skills and being able to operate autonomously.

The competency based approach has also been critiqued as not being able to capture the complexities or holistic nature of university educated professional practice. The NTB (now ANTA) has been criticised as defining competency too narrowly despite the notion of Key Competencies (Ling, 1999). This concern is certainly captured in the commentary by Geffen (1992) which emphasises concerns about translating the complexity of professional skills such as those demonstrated by health care professionals into a checklist of competencies. Furthermore, he highlights the need for education of professionals to equip them for future practice rather than immediate competency.

However, there are a number of strong arguments for suggesting that a competency based approach has relevance to professional competence and that concerns over the application of CBE concepts to university educated vocations are not well founded, and these are discussed in the next section.

2.2.2.3. Relevance of Competency Based Education to University Educated Professions

There are two main areas that require clarification when considering the relevance of CBE to university educated professions. The first relates to the types of knowledge required for competent practice of a profession and the second to the way in which competencies can be conceptualised.

Jones (2000) argues that concerns about competency based education for university educated professions are rooted in valuing a different kind of knowledge. She suggests that universities value knowledge that can be articulated, either verbally or in writing, over knowledge which is tacit or demonstrated through performance (Jones, 2000). This view is held by other authors, for example, Sefton (2001) suggests that propositional knowledge is most highly valued in academic organisations and that this is based both in history and the extent to which knowledge may be validated.

Jones (2000) also argues that both tacit and explicit (propositional) knowledge are of equal value with tacit knowledge being critical to competent practice of a vocation, including that of university educated professionals. This fits with the trend towards promoting tacit knowledge, as evidenced by performance, as a valid form of professional knowledge in the writings of many authors concerned with university based vocational education e.g. speech pathology, nursing, physiotherapy (Barrows, Williams, & Moy, 1987; Benner, 1984; Benner, Tanner, & Chesla, 1996; Best & Rose, 1996; Gamble, Chan, & Davey, 2001; Higgs, Titchen, & Neville, 2001; McAllister, 1997; McCormack & Titchen, 2001; Sefton, 2001; Titchen & Ersser, 2001). These authors argue that tacit knowledge, and personal knowledge developed by the individual through their practice, is as valid as more formal forms of academic knowledge demonstrated by performance in traditional academic assessments. CBE and assessment focuses on identifying and assessing these critical components of tacit and personal knowledge and acknowledging their critical value to competent practise of a profession.

A number of authors also argue that the Key Competencies endorsed by the Mayer Committee are also required for appropriate professional action. Down et al. (1999) suggest that the terms ‘competency’ and CBE have been interpreted in a much more narrow and behavioural sense by the university sector than the VET sector. They argue that the Key Competencies enable the transfer and application of knowledge and skills developed in educational situations to the workplace and that this is similar to the role of graduate attributes in enabling competent exercise of professional judgement and action. Down et al. also suggest

that the Key Competencies identified in the VET sector are in fact well represented in statements of graduate attributes developed by a number of Australian universities (Down et al., 1999). This notion of Key Competency is very similar to the term ‘generic competency’ defined in Section 2.2.1.5. above.

On the other hand, it is only fair to point out that commentary on research into how the Key Competencies have been applied in the VET sector suggests that it has indeed been narrowly conceptualised in that sector as well (Ling, 1999). In fact the extent to which they have been incorporated into educational programs has been termed ‘disappointing’ (Curtis & Denton, 2002).

Competent performance within the professions is also believed to rest on the exercise of expert professional judgment (Benner et al., 1996; Down & Hager, 1999; Down et al., 1999; Grant, 1999; Hager, 1999; Higgs & Bithell, 2001; Jones, 2000; Ling, 1999). Concerns about CBE expressed by authors such as Grant (1999) are based on the understanding that competencies only describe simple behavioural objectives and neglect the aspect of professional judgement and related action. Key Competencies have been proposed as underlying the competent exercise of this judgement. It seems likely that there is a continuum of ‘judgement’ along which various occupations would be placed. This continuum would span from simple task performance requiring little or no exercise of judgement through complex task performance requiring the frequent exercise of judgement in constantly changing environments.

The original Australian Standards Framework (ASF) levels did address the increasing complexity of judgement inherent in the continuum from technical to professional levels of qualification or competence, as well as the increasing degrees of autonomy and knowledge base required (Harris et al., 1995). For example, Harris et al. (1995) outline the descriptors

related to the lowest level of recognised qualification (provided through a VET provider) which require students to demonstrate the following types of competencies (pp. 58):

1. Application of knowledge and skills to a limited range of roles.
2. Choice of actions required within established routines, methods and procedures.
3. Direct guidance with regular checking or work teams.

Level 8, the highest level, and relevant to tertiary level competence required the graduate to demonstrate the following competencies (pp. 57):

1. Self-directed development and mastery of a range of knowledge and skills, applicable to broad and varied and/or highly specialised contexts.
2. Normally full independence and contexts and combinations of great variability
3. Complex judgement is applied in planning, design, technical and/or management functions.
4. Responsibility and accountability for the work of others and general functions.

This distinction is less explicit in the Australian Qualifications Framework (AQFAB, 2002), which superseded the ASF levels, but would appear to also be acknowledged in their descriptions of the key features or learning outcomes that correspond to the 12 levels of qualifications available in Australia, which span from Senior Secondary Certificate of Education through to Doctorate degrees.

In summary, CBE and assessment philosophically allows for competencies to describe the complexity of work encompassed by professional occupations, despite its origins in the 'non-professional' VET sector. Given the compelling reasons previously identified as to why competency should be assessed, defining both the generic and occupational competencies required for competent practice of a profession becomes the next challenge.

2.2.3. Defining Professional Competencies

The case has been made that competency based education and assessment can be applied to the work domains of professional occupations. It was asserted that defining competence,

which in turn informs the development of the assessment of speech pathology competence in this research undertaking, is a matter of identifying:

1. The appropriate scope of practice.
2. The generic and occupational competencies required for competent performance across this scope of practice.
3. Determining the level of performance required for a person to be deemed sufficiently competent to be qualified as a professional practitioner in that field.

It is self evident that defining competence for a particular profession will be a complex task! However, it is a key component of assessing the competency of a particular professional's performance, as identified by Benner (1984):

> "Performance measurements can be only as productive and accurate as the competencies selected to be measured. Measurement techniques, no matter how refined, cannot overcome the limitations incurred in the identification of competencies to be measured." (pp. 43)

2.2.3.1. Defining Scope of Practice

Defining the scope of practice across which a professional can claim to be competent is possibly the least troublesome, particularly for professions that have been established for some time. Historically the formation of professional associations has arisen from the need of a relevant group of practitioners to identify and defend a particular area of competence territory as theirs, ensuring that their exclusive knowledge and skills are sought by the market (Eraut, 1994; Higgs & Bithell, 2001). This group would then claim that all and only the members of their specific professional group are competent in this particular field and will regulate and protect the right to enter the profession through requirements such as particular qualifications, experience etc.

The scope of practice will naturally change over time, with new 'unoccupied' territories emerging. New areas of work can be picked up by a number of professional groups and will become part of the core competency of one professional group as a result of the current expertise, political influence, and entrepreneurial talent of the members of that group (Eraut, 1994). Given that professional territory will be well defended due to the market implications

of ‘losing’ territory, definitions of scope of practice for a profession need to be made carefully.

2.2.3.2. Defining Generic and Occupational Competencies

Defining the generic and occupational competencies for a profession is more problematic. Clearly each profession is currently practising and marketing its specific expertise and has some consensus, probably both tacit and explicitly defined, about what it is that they are qualified to do. Indeed a number of countries other than Australia have defined what competencies are expected of entry level practitioners including the United States (ASHA, 2000) and United Kingdom (QAAHE, 2001). These documents outline the occupational competencies of the speech pathology profession i.e. what a speech pathologist needs to be able to do to be considered competent when commencing practice.

On the other hand, there is much debate in the literature as to the nature of professional practice and expertise, and how it might be defined and assessed in a meaningful way. The fact that this debate is occurring suggests that it is important to examine what is understood by professional competence and how this may relate to the generic and occupational competencies required to practice a profession competently.

2.2.3.3. The Nature of Professional Practice.

Harris (1993) has identified that the epistemology that dominates when considering the nature of professional practice is that of ‘technical rationality’ where professional activity is seen as simple selection of the appropriate means to create the appropriate outcome. Harris contends that such models are simplistic, in that they assume that the means are obvious and clearly applicable, and the end is an outcome that is explicit and exists in a stable institutional context.

Professional practice is more complex than this and is characterised by a number of features. First, practice occurs in complex, changing environments where a certain degree of uncertainty is likely to exist and unique problems may be encountered. Second, it exists in a specific context which includes history, socio-political, and economic cultures (Higgs, 1999). Third, it is uniquely individual and has developmental aspects where professional practice results in the professional developing new ideas, understandings, and skills that may change who they are and what they do (Higgs, 1999).

Higgs & Bithell (2001) outline a number of models of health professional practice, education, and expertise that have influenced notions of professional competence, which are summarised in Table 1.

Table 1. Models of Health Professional Practice Summarised from Higgs & Bithell (2001)

Name of Model	Date	Description
Apprenticeship	Prior to the 20th Century	Closely guarded practice knowledge acquired on the job and expertise developed depended on the quality of the Master's competence, tuition and feedback and the learner's work based experience. Focus on practical knowledge, craft and art of the practice role of the health care worker.
Health Professional	Early 20th Century	Clinical and technical competence supported by a more scientific knowledge base, with the profession assuming responsibility for knowledge generation and quality control. Notions of professional responsibility and service quality and professional self interest developed.
Clinical Problem Solver	Mid to late 20th Century	Rediscovered the central link between reasoning/problem solving and knowledge. More emphasis on self directed learning, clinical problem solving and thinking skills. Domain specific knowledge emphasised and included in curricula and practice.
Competent Clinician	1970 – 1980	Competencies emphasised and expertise related to cost efficiency, cost effectiveness and demonstrable competencies.
Reflective Practitioner	1988	Expertise involves an advanced degree of higher level cognitive skills involving reflection on professional action, in addition to knowledge, technical ability and interpersonal competence
Scientist Practitioner	Mid to end 20th Century	Professional expertise is related to scientific rigour and evidence based practice to establish professional credibility through emphasising its scientific basis.
Interactional Professional	1999	Health professionals need to be client centred and equipped with generic skills to enable ongoing professional development and substantiation of their practice along with effective interaction with the context in which they are working.

The above models are seen to reflect trends in the interlinked areas of professional socialisation or understanding of health professional practice, education, and understanding of what constitutes expertise (Higgs & Bithell, 2001). An educational paradigm that has also strongly influenced thinking about what is involved in competent professional practise is the taxonomy of learning outcomes developed by a committee of colleges led by Benjamin Bloom (Bloom, 1994).

This taxonomy was originally developed to guide assessment of learning outcomes within tertiary programs and described a developmental sequence of learning objectives to be applied across three domains of learning. However it has, since its original development in the 1950's, been applied much more broadly. The three domains described were cognitive, affective, and psychomotor aspects of the learning. These domains and the hierarchy of learning objectives within these domains is commonly referred to as Bloom's Taxonomy.

Cognitive aspects include knowledge and development of intellectual skills including recall or recognition of specific facts, procedural patterns, and concepts that enable the development of intellectual abilities and skills (Clark, 1999). The categories composing the cognitive domain are structured in a developmental order and are thought to reflect a progression from the simplest behaviour to the most complex: knowledge; comprehension; application; analysis; synthesis; and evaluation (Bloom, 1994). This domain is commonly referred to as the 'knowledge' domain (Bloom, 1994; Clark, 1999).

Affective aspects include feelings, values, motivations, and attitudes. It has five major categories, which also operate in a developmental progression: receiving phenomena, responding to phenomena, valuing, organising and internalising values (Clark, 1999). This domain is commonly referred to the 'attitude' domain, although Carter (1985) calls it the 'feeling' domain.

Psychomotor aspects include physical movement, coordination and use of motor skills, which develop with practice and are measured according to speed, precision, distance, procedures or techniques in execution. A number of categorisations were developed to describe the psychomotor domain from 1966 to 1973 with none being published as the proposed third and final handbook in the Taxonomy of Educational Objectives series edited by Bloom and others (Krathwohl, 1994). Krathwohl and Stewart's classification (Krathwohl, 1994) focussed on physical movement and identified seven major categories: readiness, movement skill development, movement pattern development, adapting and originating movement patterns. The domain is commonly referred to as the 'skills' domain (Clark, 1999) but perhaps is better described as being about 'doing' (Carter, 1985). This notion of skilled behaviour being underpinned by relevant knowledge, skills, and attitudes has strongly influenced the practical components of health professional programs in Australia (e.g. Best and Rose, 1996) and is also referred to frequently in the literature on professional competencies.

2.2.3.4. Knowledge, Skills, and Attitudes in Professional Practice

How do current researchers and authors see the concept of knowledge, skills, and attitudes applying to professional practice? It would appear that in practice the use of the knowledge/skill/attitude classification has diverged markedly from the categories represented in the original Bloom's taxonomy.

Knowledge

Knowledge in the professions has generally been conceptualised as theoretical knowledge that can be articulated (Jones, 2000), reflecting the tertiary-educated origins of the professions. However, ultimately professional education aims to produce professional graduates with practical skills and the nature of practical professional knowledge has been the focus of much debate. Hager (2000) terms this as 'know-how' or the type of knowing what to do in practice and suggests that it is not well understood. This is different from 'know-that' or theoretical knowledge and it is suggested that it is acquired independently of theoretical knowledge (Benner, 1984).

Beeston & Higgs (2001) suggests that there are three major kinds of professional knowledge that the novice practitioner must possess. First, propositional knowledge or 'know-that' which is associated with practice and is based on theory and research. This is the type of knowledge that tends to be highly valued in the 'technical rationality' and scientist practitioner models.

The second type of knowledge is related to practical professional knowledge or 'know-how' which Beeston and Higgs term 'professional craft knowledge'. This knowledge is developed through professional experience and enables students to develop the skills essential for the practice of their profession. Such 'craft' knowledge is often tacit and sometimes intuitive, and guides day to day action by the professional enabling a rapid and fluent response to situations (Titchen & Ersser, 2001). Many authors suggest that such knowledge cannot be articulated, or at least not entirely e.g. Epstein & Hundert (2002), Jones (2000). Others insist that this does not mean that it is entirely impossible and should be attempted (Benner et al., 1996; Eraut, 1994; Hager, 2000; Harris, 1993; Titchen & Ersser, 2001). Indeed, knowledge made explicit is more easily learned (Hager, 2000).

The third kind of knowledge that a practitioner will possess is personal or experiential knowledge. This is the knowledge that each individual brings to whatever they do and is

based on life experience and accrued unconsciously (Beeston & Higgs, 2001; Epstein & Hundert, 2002; Higgs et al., 2001). Interestingly this type of knowledge has significant overlap with the affective domain of the Bloom's taxonomy as it includes beliefs and values that provide a frame of reference for the individual to act and engage with colleagues and clients. In addition, the cognitive component of Bloom's taxonomy appears to be most useful in dealing with propositional knowledge and not professional craft and personal components of knowledge.

This interaction and overlap is also noted by Harris (1993) who comments that a lot of professional practice requires a convergence of multiple sources of knowledge to enable the professional to make judgements and take wise action. This process is necessary to deal with the conflicting values and ambiguous outcomes that exist in professional practice. This is similar to the notion of professional artistry that can be understood as a further development of practical or craft knowledge with experience. Beeston & Higgs (2001) suggest that this requires development of all three kinds of knowledge in a synergistic fashion. Artistry can be seen as methods that are extremely idiosyncratic and are characterised by originality and evidence of invention (Harris, 1993). This degree of expertise will not of course be expected for the entry-level practitioner.

Skill

The range of skills considered to be an essential part of a particular professional practice is broad. There are certainly skills that could be classified as occupational competencies and directly relate to the competent practice of the profession. However, as identified by Schuwirth et al. (2002), capturing the complexity of professional practice is not straightforward and identifying the parts is not sufficient to describe the whole.

Harris (1993) suggests that there are three components of skilled professional practice: technologies, craft, and art. Technologies are the most easily articulated aspects of professional skill as they are established, prescriptive, have to be followed exactly, and have a particular type of result in mind e.g. some assessment procedures in speech pathology. These are the skills that are most likely to be addressed when assessing competency as they are the most easily defined (Chapman, 1998).

'Craft' skills are part of established practice but may be modified uniquely by the individual (Harris, 1993). This implies that the individual will need to exercise some discretion about how the technique is used, and it may be applied to a wide range of

circumstances that are aimed at a more indefinite result e.g. improved communication outcome for the client. The concept of 'artistic' skill in relation to professional practice, as is 'artistic' knowledge, would not be expected of an entry-level professional.

Another skill that receives a lot of attention in literature on professional competence is reflection to assist in articulating, reviewing, and developing the professional's understanding of why and how they do something and what influences their practice (Hager, 2000; Hays, Jolly et al., 2002; Higgs & Hunt, 1999). This skill is seen as critical for the development and application of all types of knowledge.

Schon, a key writer in the area of reflective practice and its relationship to the development of knowledge for professional practice, highlighted the notion of 'knowing in action' that relates to the concepts of professional craft knowledge and skill (Schon, 1987). Schon also highlighted the importance of critical reflection during and after action as an important source of knowledge for the practitioner and an important component in developing professional skill.

The notion of 'judgement' as an essential professional skill has also received attention in the literature on professional competency. Hager (2000) suggests that practical judgement is an essential component of workplace competency and that this is a mental skill that can be taught. Judgement involves discovering relationships between phenomena and making a judgement on these relationships. Hager (2000) describes three orders of judgements starting with generic judgements that are abstract and involve identifying similarities and differences. The second order of judgements are mediating judgements that evaluate causation, value, fact, and relevance. Finally, culminating judgements are exercised, these are the least abstract and apply the judgement to the practical situation e.g. social, ethical, scientific, technological, professional, and aesthetic judgements.

Hager (2000) suggests that a reflective model of education should encourage the development of all three orders of the skill of judgement, otherwise incorrect or poor culminating judgements will occur. The appropriate exercise of the skill of professional judgement will enable the professional to respond appropriately to the ever changing and complex context of the workplace environment.

Benner et al. (1996) highlight the aspects that influence mediating judgements during the practice of clinical judgement in nursing. They suggest that there are five interrelated aspects of clinical judgement that characterise the responses of expert nurses in practical workplace situations. These include considering underlying moral dimensions i.e. disposition toward

what is good and right, relying on extensive practical knowledge developed from previous experience, and using intuition or immediate apprehension of a clinical situation that is then subjected to deliberation to evaluate it. The influence of the context of the situation and the nurse's personal reactions to these are also attended to and a narrative is engaged that helps direct attention to the human motives, intents, and meanings as well as the biological aspects of the disease. Thus it would appear that the skill of judgement cannot operate independently of professionals' knowledge base and attitudes.

It is suggested that there are other more general or generic skills that are considered to be part of professional practice. For example, it is proposed that the community expects health science graduates to have the following generic skills: analytical; critical thinking; problem solving; lifelong learning; and communication skills (Hunt, Adamson, & Harris, 1999). Many such lists exist and these skills could be seen as being based on the skills of judgement and critical reflection; as well as being related to the category of professional craft skills. It may be that each profession values or emphasises particular kinds of skills but all are likely to require graduates entering the profession to have the skill of being able to maintain and develop their competency.

Overall, it would appear that the Bloom's taxonomy conceptualising of skills as relating to psychomotor skills or physical skills only (Bloom, 1994) may not be a satisfactory classification in practice. It appears that the literature sees many of the cognitive aspects of the taxonomy as being a skill rather than a knowledge.

Attitudes

The literature abounds with examples of the types of attitudes needed for appropriate professional action and once again it is likely that each profession works from its own particular value base. Hunt & Higgs (1999) suggest that many of the generic skills that are identified as desirable graduate attributes by universities are in fact more reflective of personal values and attitudes, rather than skills. It is suggested that professionalism assumes an underlying value system that includes a willingness to be accountable, to recognise limitations, to be tolerant, to have integrity, and apply professional expertise in a way that considers the good of the environment and community (Hunt & Higgs, 1999).

The practice of a profession cannot and should not be separated from the values and attitudes of the person and their frame of reference (Higgs, 1999) and thus is closely related to their personal knowledge. This is also pertinent to the exercise of appropriate ethical action.

Critical attitudes identified for the ethical practice of medicine have included: honesty; integrity and trustworthiness; empathy and compassion; respect and responsibility; a willingness to critically self appraise and to pursue a course of lifelong learning (ATEAM, 2001).

2.2.3.5. Competencies and their Relationship to Knowledge, Skills, and Attitudes

It is already apparent from the above discussion that the knowledges, skills, and attitudes that are considered important for the competent practice of a profession do not exist independently of each other. Therefore this must be acknowledged in the development of an assessment instrument. An example of this interrelatedness is the practice of clinical reasoning in the health professions. Medical educators have long identified that clinical reasoning skills cannot exist independently of a sound knowledge base and that skills in both these areas develop at the same time (Boshuizen & Schmidt, 2000). Allied health educators have also identified that not only do clinical reasoning skills require this knowledge base but also an ethical awareness to inform the critical thinking processes involved (Ferguson, Gibbons, Van Der Wal, James, & Baines, 2001).

As another example of this interrelatedness, Dreyfus and Dreyfus (1996) suggest that practice without theory cannot produce fully skilled behaviours in complex situations and that theory without practice has even less chance of success. They propose that theory and practice, including skills of intuition and judgement, operate in a ‘bootstrapping’ process where one supports and promotes the other as skills are developed (Dreyfus & Dreyfus, 1996).

This of course further complicates the decision as to what the essential components of competency might be, but reinforces the previously identified idea that competency is more than the sum of the parts and is a complex, interrelated phenomenon. However, it may be that the generic competencies identified and valued by health professions as fundamental to competent practice are different from the seven Key Competencies identified by the Mayer report.

Thus it can be argued that the generic and occupational competencies of a profession will arise from an interaction of the knowledge, skills, or attitudes required to carry out these tasks competently. This clearly is in agreement with the viewpoint described previously that generic and occupational competencies are intertwined in practice and are combined to specify the knowledge, skills, and attitudes necessary for competence. Thus when the components of

professional activity are analysed, they represent a combination of occupational and generic competencies and particular aspects of the work context.

However, it is worth considering whether there are other models of professional competency or education that would also contribute to a clearer understanding of the nature of professional competency.

2.2.3.6. Models of Professional Competency

Models of Learning Outcomes for Professional Education

It would appear from the previous discussion that Bloom's taxonomy, while describing important categories commonly understood as knowledge, skills, and attitudes, does not satisfactorily describe the learning outcomes a graduate professional should have achieved on entering their profession. Bloom's taxonomy is not the only model of learning outcomes relevant for professional education. Carter (1985) published a taxonomy of objectives for professional education that he felt addressed some of the shortcomings of other models. He critiques Bloom's taxonomy as not distinguishing between knowing how to do something and actually being able to do it, and as being more suitable for defining objectives at a detailed level rather than at the macro curriculum design level. The categories are also criticised as being too broad but at the same time not inclusive of some topics that should be included. Carter (1985) suggests that Bloom's taxonomy does not apply well to professional education where it is necessary to be very specific about the qualities students should have and this may include qualities that are not usually included in a typical academic curriculum, which the taxonomy was designed to address.

Carter considers his taxonomy to express the qualities the students should posses to be able to practice as professionals as opposed to the learning experiences that will develop these qualities. His taxonomy has three principle divisions that have two major subgroups of cognitive and affective components. The first division is that of knowledge or what students know and includes factual (cognitive) and experiential (affective) knowledge. Second is skill, or what students can do which includes the cognitive aspects of mental and information skills, and affective aspects such as action and social skills. The third division is personal qualities or what the students are, with cognitive aspects including mental characteristics, attitudes, and values as well as affective aspects including personality characteristics and spiritual qualities.

It is interesting that this taxonomy would appear to owe a debt to Bloom's taxonomy as its major subdivisions correspond somewhat to the categories of cognitive, affective, and psychomotor. However, Carter's taxonomy is more helpful than Bloom's taxonomy when considering professional education due to its emphasis on the qualities required for professional practice rather than the types and progression through levels of learning that needs to occur.

Carter's taxonomy is a more holistic representation of what qualities a professional requires for effective practice and acknowledges that the relative importance of each category will depend on the profession being analysed. The taxonomy is sufficiently open ended to allow for each profession to tailor the content of each category for their particular practice. For example, under 'personal qualities' engineers would be expected to place significant emphasis on the value of 'things' as it is important that they believe in the value of the products they are helping to make. Teachers however, need to place emphasis on the value of each person as an individual and managers need to value the contributions of groups of people to the success of their company (Carter, 1985). Overall the 'personal qualities' component of the taxonomy allows for a broader conceptualising of the affective nature of professional practice.

The 'knowledge' component of the taxonomy values both factual knowledge beyond the acquisition of mere facts but also including structures, procedures, concepts, and principles. Furthermore the value of experiential knowledge is acknowledged through including this as a significant component of professional knowledge. Overall, knowledge can clearly be seen as only one component of three required for appropriate professional practice. This places the concept of 'knowledge' in an appropriate context for successful professional practice. Personal knowledge, while not included in the 'knowledge' section of the taxonomy, is represented under 'personal qualities'.

Finally, the 'skill' category of the taxonomy provides a useful breakdown of the types of skills required, without undue emphasis on manual or technical skills, and can be seen to subsume mental skills such as judgement and reflection. This allows for skills to be considered that are more than purely psychomotor or movement abilities and that could be applied across the knowledge and personal qualities components of Carter's taxonomy.

Carter's model was primarily developed to ensure that the university curriculum is aligned with what the profession requires for competent practice and thus has included consideration of what the components required for professional competency are. Three other models exist in

the literature relevant to the competent practice of health professionals, or more specifically, speech pathology.

Models of Professional Competency

Miller's pyramid (Miller, 1990) has had significant influence upon the conceptualising of competence in the medical education field and also health profession education generally. This model has already been described earlier in this literature review as well as how this relates to notions of competence underlying performance. Miller and more recent writers such Schuwirth et al. (2002) identify that the 'knows' and 'knows how' layers are the easiest to assess but that these cannot be assumed to predict fully and with confidence the students' ability to perform competently in actual practice. All acknowledge that accurate and reliable assessment of this level of the pyramid is difficult. This model also does not provide a blueprint for identifying what the generic and occupational competencies for successful performance of the profession might be.

Further to this, Rethans et al. (2002) identified that Miller's model does not account for the influence of other factors on clinical performance, both external or systems related issues (arising from the context of practice) and internal or individual related influences (associated with the individual's state of mind and personal context). Their 'Cambridge Model for Performance and Competence' is a refinement of Miller's model and suggests that competence is an important pre-requisite for performance and that both are illuminated by these two other groups of factors such that not all problems with performance can be attributed solely to the practitioner's underlying competence.

Another model of competence in health professionals is "The Interactional Professional" developed by Higgs and Hunt (1999). This model clearly identifies the need for competence to relate broadly to the context in which the professional operates. They suggest that there are four broad areas of professional competence. These areas include competencies that are more obviously linked to the interaction with the client and operating effectively within the workplace such as technical competence, which is both generic and discipline specific and interpersonal competence. However, they include two competencies that reflect a broader view of the context of professional action. This includes the ability to interact with and change the context of practice, relating to being problem solvers and change agents, and a professional responsibility to serve and enhance society.

With this model Higgs and Hunt (1999) are attempting to highlight that professional practice should not be judged only on the basis of specific professional skills but also on the more generic ability to effectively interact with the context of practice. This includes being able to work with a range of clients and colleagues and making decisions in different settings within a changing environment. Ethical notions such as professional responsibility to society and accountability are central. Affective components of professional practice are also identified such as sensitivity, respect for, and empowerment of clients. This type of model acknowledges that the professional operates in a broader context and that both generic and occupational competencies are required to successfully operate in this context.

A final way of conceptualising competent practice that is related to the practice of speech pathology, is not so much a model as a metaphor, termed the 'Iceberg Metaphor' (Fish & Coles, 1998).This metaphor was proposed as a way of conceptualising or illuminating the nature of the practice of speech pathology. Fish and Coles (1999) suggest that the 'doing' or performance of speech pathology practice is the 'tip' of the iceberg and is what can be seen.

However this performance is built on underlying components, of which there may be little or no awareness. How the professional experiences practice is seen to be at the 'waterline' of the iceberg, reflecting some awareness of what the professional is doing and saying. However all this is underpinned by less conscious knowledge (both personal and formal) as well as assumptions and expectations that are bedded onto personal attitudes, beliefs, and values. In line with the iceberg metaphor it is suggested that if 'doing' is given undue emphasis, and no time is devoted to what exists under the waterline, the iceberg may capsize. Again, this metaphor draws upon ideas of underpinning or generic qualities including different types of knowledge, skills, and personal qualities.

It is apparent all of these models identify, to a greater or lesser extent, competence that arises from both the professional's generic and occupational competencies and the way in which these competencies are applied to specific professional contexts. In addition, the role of the context and its influence upon the professional's competent performance is also acknowledged.

2.2.3.7. Defining Professional Generic and Occupational Competencies: Summary

It has been proposed that the CBE literature accepts that there are generic competencies and occupational competencies that combine to create competent occupational performance (See Section 2.2.1.). The generic competencies are considered to be underlying combinations

of knowledge, skills, and personal qualities that contribute to competent performance; such as the seven “Key Competencies” have been accepted by CBE. Occupational competencies are considered to be occupation specific skills that include varying combinations of aspects of the generic competencies, including knowledge and attitude components.

The literature on professional competency would appear to endorse the concept that there are both generic competencies and specific occupational competencies required for the competent practice of a profession. There appears to be consensus that there are three important components that interact to create both the generic and occupationally specific competencies: knowledge, skills, and personal qualities. Knowledge refers to propositional, practical (professional craft knowledge), and personal knowledges. Skills may include practical skills such as technologies and the individual application of professional ‘craft’ knowledge, cognitive skills such as reflection and judgement, and possibly others. Personal qualities encompass the unique mental characteristics, attitudes and values, personality characteristics, and spiritual qualities of the individual.

Thus competent performance appears to be the end result of a complex interrelated web where knowledge, skills, and personal qualities (KSP) combine in various ways to create generic competencies (that are represented in various combinations within occupational competencies). In turn, these generic competencies make competent performance of specific occupational tasks (or occupational competencies) possible both in the present and into the future.

Nonetheless, it is evident that any statements of professional competency are likely to be outlines of the occupational competencies required for competent performance of that particular professional role. It can be predicted that these occupational competencies will be derived from the interaction of various generic competencies, along with different types of knowledges, skills, and personal qualities.

At the start of this section on defining professional competencies it was identified that competence needs to be defined with regard to scope of practice, generic and occupational competencies required across this scope of practice, and with regard to the level of performance required to be deemed sufficiently competent to enter the practice of speech pathology. The discussion to this point has primarily elaborated on what is meant by generic and occupational competencies in general terms. The next section will identify what information on speech pathology competencies currently exists, particularly within the Australian context.

All professions conduct some sort of evaluation of competence before allowing admittance to their professional group and indeed this research endeavours to enhance the validity of this process. This suggests that all professions have some concept of what comprises a sufficient level of performance for 'entry-level' membership. However, it is also reasonable to assume that the degree to which someone is competent will change over time and exists on a continuum such that the standard of performance deemed to be competent is a particular point on this continuum (Ling, 1999). The significant issues then for assessment are: what are the characteristics of this continuum and what is the 'point' that indicates competent performance? These issues will be addressed in Chapter Five when considering assessment tool design.

2.2.4. Speech Pathology Competencies

2.2.4.1. Specifying competencies

How should competencies be specified for a particular profession? Research suggests that field educators or practitioners value different aspects of competencies than do university educators (Cross, 1998; Loomis, 1985b; Neary, 2000a). For example, Cross (1998) found that from eight dimensions of physiotherapy competency, CEs identified communication, disposition, and commitment as being relatively more important. Academics, on the other hand, saw the competencies of knowledge, approach to learning, and commitment as paramount. Cross suggests that both groups are focussing on generic rather than profession specific skills but she suggest that academics have a longer term focus than 'fitting in' in the workplace.

Another study found that a medical education faculty placed greater importance on medical students' skills in organising and applying information than on their skills in interviewing a patient, correctly performing a physical examination or relating to a patient (Solomon, Speer, Callaway, & Ainsworth, 1996). It is possible that, in the first instance, students and possibly their field educators, will be most concerned that these technical competencies are developed.

Loomis (1985) found that physiotherapy practitioners, as opposed to academics, appeared to place more emphasis on technical knowledge and skills rather than affective components of competence such as communication with patient and family members. She suggests that caution should be used when competencies are compiled from what practitioners believe to be

important as they tend to relate to their own, necessarily idiosyncratic, professional practice and this creates wide variation in ratings of what competencies are important.

Similarly, Milton (1999) suggests that it is important for university programs to work from a well thought through view of professional practice based on a rigorous analysis of this practice. He argues that the way the profession defines competence will structure or form their knowledge and skills; and that the professional curriculum should focus on learning objectives that are relevant to the profession. Unlike Milton (1999), whose suggested process does not appear to involve actual consultation or observation of professional practice in real contexts, Benner (1984) states that it is important to identify the competencies evident in actual clinical practice. This has rarely been done in speech pathology and given the argument that much of health professional practice is based on unarticulated craft knowledge (Titchen, 2001), this approach appears to have significant merit.

One example is the research conducted by Ferguson & Elliot (2001) who looked at the process of therapeutic interaction in aphasia therapy. They noted that there is very little guidance in the literature for students about how such sessions are conducted and that their analysis helps make explicit the implicit process of therapy. However, it can be equally argued that this type of focus may hinder the development of a profession through focussing on the practical competencies needed 'today' and neglecting those needed for 'tomorrow' (Ling, 1999).

It has been suggested that statements of competence need to be realistic, general, and embody requirements of good practice and not actual practice (Miller et al., 1998). It is these differences that can create stress for students, as they live with the difference between practice reality and university ideals and the way in which field educators handle these (Neary, 2000a). Developing realistic, general statements of good practice would require negotiation between practitioners, managers, professional bodies, recent graduates, and the academic programs. Carter (1985), as did Robertson, Simons, & Harris (2000), identified that these groups may have quite different perceptions of the demands of a profession but optimistically suggests that information from all these sources can be compared and a consensus arrived at that would create a starting point for curriculum design.

2.2.4.2. Australian Statement of Speech Pathology Competencies (CBOS)

As previously mentioned, speech pathology in Australia has developed a statement of the competencies required for entry-level practitioners in its CBOS document (SPAA, 2001) in

response to the need to assess the competency of overseas qualified speech pathologists. This document "outlines the **minimum skill, knowledge base** and **attitudes** required for **entry-level** practice of the profession." (pp. 1, author's own bolding). As such it endeavours to outline the occupational competencies required to be sufficiently competent to enter or commence practice of the profession, the scope of practice for these competencies, and the level to which they should be performed[3].

The process used to develop the CBOS was fourfold (Dawson, 1993). A Steering Committee oversaw the project and comprised five members of the professional association, one representative respectively from the universities, unions, employers, and the federal government (NOOSR). A reference group was drawn nationally from the profession, representing a wide variety of practice in speech pathology. This group worked together to formulate the core competencies. Telephone interviews with new graduates were conducted using a critical incident technique to identify the different attributes in practice that were to be reflected in the units of competency. A cycle of drafts and revisions was undertaken, with feedback sought from the profession and consultations with each state's professional association and university program(s).

This process resulted in a CBOS document for entry level standards for speech pathologists in May 1993. The document was subsequently revised in 2001. The format of the CBOS conforms to the accepted CBE practice in Australia in that it uses the following structure, starting with a Unit level heading (SPAA, 2001):

1. Units: this heading and its descriptive paragraph outline a broad area of professional activity.
2. Elements: specific activities carried out within the specified unit.
3. Performance criteria: examples of behaviours that would indicate that the elements of competency are being carried out to an acceptable standard.
4. Cues: illustrations of the knowledge base; practical considerations; actions; attitudes; and some contextual features that are evidence that a performance criterion has been achieved.

The CBOS also includes a range indicator statement specifying the range of ages and areas of practice a speech pathologist must be able to cover as well as the level of independence required for competent performance at entry level.

[3] The CBOS document is available from here: http://www.speechpathologyaustralia.org.au/Content.aspx?p=78

The CBOS statement identifies that there are three interrelated components for successful practice: knowledge; integration and application of this knowledge and occupational skills; and ethics. The CBOS focuses on detailing the occupational competencies or behaviours required to practice the profession, rather than underlying generic competencies, knowledge, skills, or personal qualities. Thus generic competencies and the related KSP are inferred from observing competent performance of the occupational competencies. However, some of the relevant KSP are identified at the cue level of the CBOS and the need to consider the CBOS statement as an integrated whole is emphasised. In addition, a set of range indicators is provided to describe the scope of speech pathology practice.

Given that the CBOS is a comprehensive statement of speech pathology competencies for entry-level practice in Australia, this document will naturally be pivotal in the development of a competency based assessment of speech pathology students. However, based on the issues identified in the literature regarding specifying competencies, there are a number of issues that arise when considering the use of the CBOS competencies in the development of an assessment tool.

2.2.4.3. CBOS and Specifying Competencies

A number of questions arise when considering the development and use of entry level professional competencies described in the CBOS as a basis for assessment of competency in speech pathology practice. First, do the competencies adequately identify the complex interrelationships of KSP that are required for the professional practice of speech pathology as well as the need to be competent today and in the future? This is a question that will need careful consideration when designing the assessment tool.

Other questions that merit consideration include how has the process of developing the competencies impacted or constrained the competencies identified? Does developing competencies through asking the profession to reflect upon and articulate what it is that they do and value tend to create competencies that are about what the profession ‘thinks’ it does, rather than what it ‘actually’ does? Should competencies only be developed on the basis of actual practice? And finally, is it actually ‘good practice’ as opposed to actual practice that is described?

It is not possible to definitively answer all of these questions and unlikely that the resources would be available to enable a profession to fully respond to all the issues raised

above. However, it is useful to scrutinise the development process of the CBOS so that an understanding can be gained of its strengths and weaknesses.

Certainly the developmental process for CBOS is solely based on the profession's reflection on what it is that they think they do. This is unsurprising given the paucity of literature for most professions on actual clinical practice and the size of the undertaking to cover the breadth of practice inherent in speech pathology. However, it does create the risk that the competencies will reflect what the profession subjectively perceives itself as doing. Thus, it is important to keep in mind that there may be unidentified unarticulated knowledge, skills, and personal qualities that are important to competent entry level practice in the profession.

With regard to whether the CBOS entry-level competencies describe 'good practice' as opposed to 'actual practice', it is difficult to determine this, particularly in the virtual absence of any studies of real life practice in speech pathology. The CBOS aims to outline "the minimum skill, knowledge base and attitudes required for entry-level practice of the profession" (pp.4) (SPAA, 2001). However, this notion of 'minimum' does not appear to be differentiated from 'good practice' in that page 2 of the document states that "Speech pathologists undertake to provide a high quality service …"(SPAA, 2001). Thus it would appear that the minimum aspired to is in fact 'good practice'. However, actual practice is unlikely to consist of good practice on every single occasion as human beings are fallible. For example, personal factors may temporarily affect the speech pathologist's ability to attend to and synthesise all the relevant features of the client and work setting that influence their decision making.

This is an important consideration when assessing practice: just how consistently should the entry-level practitioner meet the standards outlined? If they are in fact 'minimum standards' it can only be assumed that the entry level practitioner should consistently perform to this minimum. However, when reviewing the CBOS standards it seems unlikely that someone would consistently meet all of them all of the time, although every practitioner would certainly aim to meet them.

For example, all speech pathologists would aspire to being able to justify the choice of management options for a client on the basis of a critical evaluation of current literature and research (Element 2.5, pp. 8, SPAA [2001]). However, it would not be realistic to assume that all speech pathologists would be up to date in their reading and evaluation of current literature at every single moment of making a management decision. The potentially aspirational nature

of the CBOS entry-level competencies will require careful consideration during the judgment process of assessment.

On the other hand, it can be argued that the competencies described in the CBOS do reflect both occupational competencies and generic competencies required for entry level practice and relevant knowledge, skills, and personal attributes required for these. The sequencing of the document describes the units of occupational competencies required, for example, assessment, analysis and interpretation, planning, and so forth. When the performance criteria and cues for each of these areas are examined it becomes apparent that a range of skills that could be argued to be generic are frequently represented, for example: critical appraisal; consultation; synthesis; investigation; clinical reasoning. Such skills are also likely to underpin a lifelong learning orientation necessary for future competent practice. In addition, a unit termed "Professional Development" is included and identifies competencies required for ongoing development of competency that would enable the entry-level practitioner to respond to a changing workplace.

The developmental process used to identify and document entry level competencies for the speech pathology profession has promoted this synthesis of occupational and generic competencies. First a wide range of stakeholders were involved in the developmental process. Membership on the steering committee for the CBOS project included universities, unions, employers, and the federal government, as well as representatives from the profession. The only primary stakeholder group not represented was the client and/or carers. The reference group specifically developing the competencies primarily included members from the speech pathology profession representing a wide range of practice and geographical areas. The competencies were also subjected to several cycles of review that enabled a broader range of stakeholders to comment. This has ensured that the resultant competencies are not likely to describe specific, idiosyncratic types of professional practices (Loomis, 1985b).

Second, the conceptual approach used to develop the CBOS document was based on the approaches recommended by Australian CBE practitioners such as Gonzci (1992). As discussed previously, the Australian approach strongly supports the notion that generic and occupational competencies are closely intertwined in practice and that successful work performance does not rest on performing discrete tasks successfully but on holistic integration and coordination of these tasks.

In summary, the CBOS entry-level competencies are based on the profession's perception of what it is that it does and the potential limitations of this approach must be acknowledged.

However, its strength also lies in the developmental process that has the potential to address both the priorities and concerns of both speech pathologists in their workplaces and the educational programs that aim to graduate competent entry level practitioners.

CHAPTER THREE

3. ASSESSING COMPETENCE

3.1. Nature of Assessment

As stated at the start of this literature review, it is clearly important to a range of stakeholders to know whether a speech pathology graduate is competent to practice as a speech pathologist. A case has been made that a competency based approach to the education and assessment of professionals is feasible and the notion of 'competency' and its components as it applies to professional work has been described. More specifically, it was identified that a process has already been undertaken by speech pathologists in Australia to develop a set of specified competencies that can be used as a basis for developing an assessment of competent performance of the profession of speech pathology.

As described in the introduction, SPAA uses this set of competencies (CBOS) as the blueprint to assess and accredit Australian university programs educating speech pathologists. Of particular concern, and the focus of this research, is how to reliably and validly assess whether students are able to competently perform in the real workplace environment provided by the practicum experience. To facilitate this process the literature was examined for information on the nature of assessment, current practices, reliability and validity issues.

It should be noted that, for the purposes of this discussion, the terms 'competent', 'competency' or 'competence' will be used in relation to performance and not as a focus of assessment in and of itself. It is performance and how to determine that performance is sufficiently competent for students to enter the profession of speech pathology that is the focus of this research. Any notions of underlying generic competence can and will only be inferred from the actual behaviours observed during assessment.

3.1.1. Purposes of Assessment

It has already been identified that there are three compelling reasons to assess the competency of speech pathology students in their work placements held by various stakeholders including the profession, the student, the recipient of services, and the employer. Nevertheless, this description of the purpose of assessment only attends to the gate keeping or regulatory nature of assessment and does not attend to the other purposes that an assessment

can serve, both intended and unintended, and need to be considered when developing an assessment.

Assessment purpose is generally categorised into two types: formative and summative. These purposes are from different perspectives, have a different locus of control, are focussed on different outcomes (Hays, Davies et al., 2002), and can be directed 'down' to the students or 'up' to the educator and/or their organisation . Summative assessment aims to be a definitive statement of the students' competence and is a primarily gate keeping function. For the provider of the education, this type of assessment may be directed to evaluating the effectiveness of the program, maintaining standards, and being accountable for the quality of teaching (Miller et al., 1998). This summative function is described by Barnard (1999) as 'bureaucratic' and related to issues of control, monitoring, and certification. Summative assessment is most obviously linked with the reasons previously identified for assessment of competency of professionals.

Formative assessment, however, is concerned primarily with assisting student learning (Boud, 2000) and is characterised by the assessment information being used to improve the performance of the student, such that the student is the central focus and participates in determining the nature of the assessment process (Brookhart, 2001). In a similar fashion formative assessment can provide the educator with feedback on their teaching (Morris, Porter, & Griffiths, 2003; Robertson, Rosenthal, & Dawson, 1997). The goal of formative assessment is to assist the student to internalise the learning targets defined by the educator, so that the student sets his/her own goals in relation to these targets and self-monitors their progress towards them (Brookhart, 2001). Thus assessment processes that include feedback to the student regarding their progress are integral to formative assessment.

Formative assessment of a particular student's ability to practise speech pathology in the workplace can include practices such as collaboratively discussing and evaluating a student's rationale for the intervention session they have just completed with a client. For example, discussing real or hypothetical cases, or the CE acting as a 'sounding board' for the student developing their understanding of professional practice (Higgs, 1997). Such processes have been influenced by the move within educational practice to view learning as an active process which includes acknowledging the influence of the students' personal understandings, biases, and perceptions (Masters, Adams, & Wilson, 1999).

When formative and summative functions are incorporated into a single assessment there are inevitable tensions between these two related but sometimes competing functions of

teaching and gate keeping (Boud, 2000). This tension is experienced by CEs as a conflict in roles between that of teacher who wishes the student(s) to succeed and assessor who must make a final decision regarding the student(s) competence (Cross et al., 2001; Duke, 1996; Ilott & Murphy, 1997). Students also experience this dilemma in that they may focus on the goal of being judged as competent and therefore be too preoccupied with how their performance may be perceived rather than the learning task at hand (Boud, 2000).

3.1.2. Assessment and Learning

Wass, van der Vleuten, Shatzer, & Jones (2001) state that "Assessment drives learning." (pp. 945). This awareness is integral to both formative and summative assessment as both are inextricably linked to the learning in which students engage. It is now a widely held understanding that assessments direct students' attention to their learning and any educator who has been asked by students "Will this be in the test?" is highly aware of this. An assessment is a statement of 'what counts' or what is valued by the educational program (Boud, 2000; Fontaine & Wilkinson, 2003) and can be argued to be the most significant influence upon the quality of student learning (Robertson et al., 1997). Learners are actively engaged in identifying and meeting the requirements of the assessment (Crossley, Humphris, & Jolly, 2002) and it is well known that students are more likely to study for the parts of their course that will be assessed (Wass et al., 2001).

Boud (2000) warns that this creates a dilemma as summative assessment negatively affects learning while at the same time attempting to measure it, placing the responsibility for judging learning with others and not the student. He argues that society is obsessed with certification, grading, public measures of performance and accountability, which relegates the focus on learning to the background and consequently the assessment processes required to promote it. This effect of testing can be complex, unrecognised, or unexpected, but can also be used in positive ways such as developing student's self assessment skills that are essential for lifelong learning (Boud, 2000).

The assessment's impact upon the learning in which students engage and the skills it may or may not promote, should also be of concern to stakeholders who wish to ensure that speech pathology students graduate competent to practice their profession including being equipped to continue to acquire and develop new knowledge. If the assessment of students' competency to practice does not assist them in developing the actual skills required for lifelong professional competence, it has in fact failed to meet its primary goal.

Formative assessment is particularly crucial in this process as students need feedback about their competence to compare with their developing understanding of what comprises competence and what they need to do to achieve competency (Brookhart, 2001). This cycle of feedback is important in the development of the students' professionalism and their confidence in themselves as a professional (Robertson et al., 1997). In addition, formative assessment provides information for students to assist them in the active process of constructing their own interpretation of what they are learning and to relate the new information provided through feedback to their existing knowledge and understandings (Masters et al., 1999).

Given that competency is a complex set of interrelated knowledges, skills, and personal qualities, it is important that the assessment promotes more than rote or surface level learning. This will be affected by students' perception of the assessment task such that, if students perceive that only surface level learning is required, they will not engage in deeper forms of learning that require understanding and manipulation of knowledge (Maclellan, 2001).

A deep level of learning also involves developing more sophisticated conceptions of speech pathology problems and the advanced problem solving strategies required for professional practise (Masters et al., 1999). If students' attention is directed by the assessment only towards acquiring specific skills, and not integration of these skills with the knowledge and personal qualities required for competent professional practice (Robertson et al., 1997), they will not be sufficiently competent to practice at entry level. Thus, assessment must promote learning that integrates speech pathology knowledges, skills, and personal qualities.

In addition, Boud (2000) suggests that assessment has an important role in providing the opportunity to develop and practice life long learning skills such as confidence in oneself as a learner, self evaluation, problem solving, and accessing learning resources. These skills are seen as critical to ensuring that students can cope with future workplace change (Boud, 2000) and, as previous discussion has identified, are critical for the effective practice of complex professional work. Robertson et al. (1997) highlight something similar when they suggest that if an assessment matches the students' perception of their performance then it gives the students confidence in their ability to self monitor. Finally, the values and attitudes of the profession will also be communicated through the assessment process (Wolf, 1995).

Thus assessment serves multiple purposes and design of assessments must attend to all of these. These considerations are summarised by Masters (1999) :

> "Methods used to assess educational achievement convey powerful messages about the kinds of learning considered worthy of recognition and reward and so are capable of influencing educational processes in positive ways by focusing effort on valued achievements and forms of learning. Equally, however, assessment methods are capable of sending unintended messages and distorting student learning if they place inappropriate weight on less important goals or address only a narrow range of valued achievements." (pp. 20)
>
> And:
>
> "The challenge in assessing educational achievement is to gather evidence of student learning in such a way that the learning process itself is supported and not undermined or distorted." (pp. 19)

3.2. Current Practice in the Assessment of Competency

There is a vast and somewhat bewildering array of different assessment approaches in the literature all professing to assess competency. This is unsurprising as competency based assessment is actually defined in very broad terms and therefore can encompass a wide range of techniques. For example, ANTA (2002) defines competency based assessment as:

> "…the process of collecting evidence and making judgements on whether competency has been achieved to confirm that an individual can perform to the standard expected in the workplace as expressed in the relevant endorsed industry/enterprise competency standards or the learning outcomes of an accredited course." (pp. 93).

A scan of health professional literature found 16 different types of assessment techniques used in the past or currently to assess students' competency to practise their profession (Table 2) and no doubt there are others that are not described in this table. All of these techniques have been used to carry out the summative or gatekeeping function of determining whether students are equipped to practise at an entry level to their profession or chosen speciality within that profession. The assessments have aimed to focus on competency to perform, rather than acquisition of knowledge alone, and have frequently been part of assessment of workplace performance or the final determination of competency to practice. So, which technique should one choose when designing an assessment of speech pathology students' competency in their practicum?

Table 2. Assessment Techniques Described in the Health Professional Literature

Technique Name	Description
Long and short cases	Observed working with an actual patient/client and assessed on aspects of the care provided. E.g. Wass et al, (2001).
Patient Management Problems (PMPs)	Written simulations with information progressively revealed as students progress through the material. E.g. McGuire (1995), Miller (1990), Newble, Norman, & van der Vleuten (2000).
Interviews/Vivas/Standardised Orals	Student answers questions related to materials/information presented to them e.g. video of a client, or presents a case. E.g. Begg & Ferguson (2004), McGuire (1995), Southgate, Cox, David, Howes et al. (2001).
Computer Assisted Simulated Encounter (CASE) or CBX project	Similar to PMPs except the computer responds differentially according to the students' responses to the case material. E.g. Edelstein, Reid, Usatine & Wilkes (2000), McGuire (1995).
Clever Robots	Computerised models. E.g. McGuire (1995).
Standardised Patient	Students carry out a task e.g. case history, with an actor trained to simulate the same patient in the same way with each student. E.g. Edelstein et al. (2000), Epstein & Hundert (2000), McGuire (1995), Miller (1990).
Objective Structured Clinical Examinations (OSCE)	Multiple timed independent stations designed to assess predetermined clinical skills. Can include specific tasks e.g. use of equipment, answering questions about findings or working with a standardised patient. E.g. Dauphinee (1995), McGuire (1995), Miller (1990).
Chart Stimulated Recall	Examiner discusses management of the patient/client according to the charts provided by the students. E.g. McGuire (1995).

Multiple Choice Tests	Paper based testing with stems of varying complexity and a choice of responses for the students to select. E.g. Epstein & Hundert (2002).
Patient assessments	Ratings from patients/clients regarding aspects of the care they have received. E.g. Ilott & Murphy (1997), Norman, Watson, Murrells, Calman, & Redfern (2002).
Peer assessments	Ratings from fellow students/colleagues regarding the assessee's competence. E.g. Epstein & Hundert (2002).
Clinical educator's observations	Assessment by supervising clinician/clinical educator or visiting assessor via checklists, rating scales or anecdotal records. E.g. Begg & Ferguson (2004), Epstein & Hunderts (2002), Fisher (1998), Miller (1990).
Concept maps	Students diagram their understanding of a case, represented as interrelated concepts. E.g. Schwabbauer (2000).
Portfolios	Students select a folio of information regarding their performance including reflective reports and examples of their work. E.g. Begg & Ferguson (2004), Schwabbauer (2000), Tracy, Marino, Richo, & Daly (2000).
Self assessment	Ratings or reflections by the students regarding their performance. E.g. Davis (2002), Schwabbauer (2000).
Purpose developed instruments	Standardised protocols e.g. The Emotional Competence Inventory – University Edition (Boyatzis & Goleman, 2001).

3.2.1. Theory Versus Technique

McGuire (1995) warns that the measurement or assessment field is technique rather than theory driven and this focus on form over substance results in an erroneous assumption that the form of an assessment determines the component of competence being assessed. Thus the issue becomes not what technique to choose but how is competency being conceptualised and what would be the most appropriate way to sample this?

It would appear that this proliferation of techniques has arisen due to both a theoretical and practical tension between an understanding of competence as 'shows how' and performance as 'does', as conceptualised by Miller's pyramid (1990). Very frequently the terminology is confused, with 'competence' being used to describe competent performance or the potential/performance divide being described as underlying 'competence' leading to

'capability' (Eraut, 1998). As outlined previously, CBE conceptualises competence as underlying competent performance and being inferred from the person's performance (Gonczi, 1992).

3.2.1.1. Assessing competence

Most of the techniques described in Table 2 above are focussed on assessing competence in the sense that the assessment is an opportunity for students to 'show how' something should be done, in a standardised and controlled environment. The standardisation is an attempt to ensure fairness so that students have an equal opportunity to demonstrate their skills and are not disadvantaged by the uncontrolled nature of the work environment. This concern is very much linked to notions of validity and reliability.

This approach is encapsulated by the Clinical Performance Examinations described earlier and advocated by Bargagliotti et al. (1999) and Luttrell et al. (1999) to avoid the 'assessment during a fire drill' effect of ward based assessment. The CBE based approach to assessment also allows for assessment to be based on 'collecting evidence' that students can demonstrate required attributes, skills, or knowledge in a variety of situations other than the workplace including portfolios or in class assessments (ANTA, 2002).

However, the relationship between competence and performance is complicated and competence is a necessary but not sufficient requirement for competent performance (Eraut, 1998; Gonczi, 1992; Rethans et al., 2002; Schuwirth et al., 2002). Schuwirth et al. (2002) define competence as how people perform in ideal conditions knowing that they are being challenged to demonstrate that they have the knowledge, skills, and attitudes required for a task. Performance, on the other hand, is described as how people behave when in real life and when they are not being observed. It is in this sense that the terms 'competence' and 'performance' will be used in this discussion.

3.2.1.2. Assessing Performance

Performance is influenced by everyday constraints including internal factors such as the professional's own health and external factors related to the context they are working in e.g. workplace constraints (Schuwirth et al., 2002). The impact of context upon practice and the fact that professional work interacts with this context at many levels (client, team, workplace, and even global) has led to descriptions of professional work as interactional (Higgs &

Edwards, 1999). The students' ability to manage these types of impacts as well as personal ones such as health and emotional issues, is seen as an important prerequisite to graduation by CEs in speech pathology (Maloney, Carmody, & Nemeth, 1997).

Techniques to assess everyday performance have generally relied on subjective evaluations in the workplace by supervising clinicians, self, and peer assessments. This reliance on workplace based evaluations, often by the same person who is responsible for teaching the student(s) on a day to day basis, is most apparent in literature related to nursing e.g.(Peters et al., 2001); occupational therapy (Ilott & Murphy, 1997); physiotherapy (Alexander, 1996; Hrachovy et al., 2000; Roach et al., 2002); sonography (Fisher, 1998); and speech pathology (Best & Rose, 1996). However, examples of similar approaches are also found in the VET sector (Curtis & Denton, 2002) and medicine (Fontaine & Wilkinson, 2003; Page, 2004). These strategies are bedevilled with criticisms regarding their validity and reliability such as their vulnerability to subjective bias, poor inter rater agreement, lack of generalisability across cases or workplaces, and so forth, by authors such as Epstein & Hundert (2002) and Dauphinee (1995), among others.

There is of course a dilemma inherent in the definitions of 'competence' and 'performance' described above when considering assessment methods. First, being assessed for competence for some candidates immediately means that the conditions are less than ideal due to internal factors such as anxiety arising from exam stress. Second, every student on placement knows that they will be assessed and this will naturally influence their performance, even on indirectly observed skills such as record keeping.

3.2.2. Case for Workplace Assessment

Ultimately, as McGuire (1995) stated, assessment needs to be driven by theoretical understandings, in this case an understanding of what comprises underlying competence and performance at a level sufficient for entry into the profession of speech pathology. The previous section on competency has clearly outlined that entry level speech pathology performance can be seen as the end result of an interaction of knowledge, skills, and personal qualities that create occupational skills as well generic and transferable competencies to ensure the ongoing competent practise of speech pathology. The case was also made that all of these components will be either observed or inferred from observed performance. This is in line with the understanding that "Professional competence is more than a demonstration of isolated competencies …" (pp. 227) (Epstein & Hundert, 2002).

Clearly the pre entry speech pathology curriculum must enable students to acquire the knowledge and skills they require to practice in their profession and ideally this should be assessed in a coordinated way that ensures the students' attention is directed to all the KSPs required by the profession. However, the ultimate goal of pre entry professional education is to prepare students for and ensure that they are equipped for current and future competent practice in the workplace. Thus it is somewhat self evident that competency assessment must include a workplace component and focus on performance in real work environments rather than entirely relying on assessments that aim to measure the development of underlying competence on the assumption that there is a direct link to performance.

Assessment of workplace performance in practica constitutes current practice in all speech pathology education in Australia, with some programs incorporating other types of assessments such as viva examinations and portfolios. This practice, in speech pathology education, of emphasising workplace assessment is now being pursued by medical education. Medical education has traditionally relied on a variety of assessments in standardised and controlled environments such as OSCEs (as do other health professions) but is now directing attention to developing assessments of performance in the workplace.

This movement has occurred due to the recognition that competence is a necessary but insufficient requirement for appropriate performance (Rethans et al., 2002; Schuwirth et al., 2002). In fact, Page (2004) states that medical education needs to re-emphasise performance based assessments in clinical contexts and that these should be an integral component of the curriculum. This is supported by Dauphinee's (1995) suggestion that future approaches in assessment of workplace competence in medicine need to focus on more frequent direct observation of students in their placements and overcome the problems inherent in the poor measurement properties of rating scales. Cox (2000) also recommends that assessment occurs in the workplace and that the focus needs to be on how to record performance in terms of capabilities in a way that captures the important qualitative detail of the performance.

However, the practicality of a focus on workplace performance assessment does not dispense with need to be theory rather than technique driven. As McGuire states (1995), in the field of medicine, there is a need to develop an acceptable theory or unifying conception of competence on which assessment and teaching can be based. This is no less true of the field of speech pathology and is addressed in the tool development phase of this research.

Concerns regarding valid and reliable measurement of students' competencies when assessing their workplace performance led in the past to medical education emphasising more

standardised and controlled assessments and other professions have followed suit e.g. nursing (Govaerts, van der Vleuten, & Schuwirth, 2002). Validity and reliability are naturally key to any assessment process without them. everyone concerned with ensuring the competence of entry level speech pathologists, cannot have confidence in assessment results provided by any tool.

3.3. Reliability and Validity

3.3.1. Validity

The concept of validity is integral to any field of enquiry and when applied to assessment or testing, validity can be usefully described as follows:

> "Test validity refers to the degree with which the inferences based on test scores are meaningful, useful, and appropriate. Thus test validity is a characteristic of a test when it is administered to a particular population. Validating a test refers to accumulating empirical data and logical arguments to show that the inferences are indeed appropriate." (para. 1, Brualdi, 1999)

The way in which validity has been conceptualised in relation to assessment and performance assessment has been subject of debate, particularly in the 1990's. This debate has shaped an understanding of validity that has evolved somewhat from the original definitions but continues to be used in some assessment literature, including in the health sciences.

3.3.1.1. Traditional Validity Concepts

Brualdi (1999) summarised the debate as moving from a traditional concept of validity where validity evidence was grouped into three categories (content-related, criterion-related, and construct related) to a more modern, unified concept of validity with six interrelated aspects. The three traditional validity categories are intended to be a way to organise and discuss validity evidence rather than clearly distinct types of validity although they have frequently been applied as if they were separate and interchangeable types of validity types (Messick, 1996). Thus evidence may fall into more than one type of validity category. The traditional categories are described in the Table 3.

Table 3. Traditional Concepts of Assessment Validity Summarised from Brualdi (1999)

Category of validity evidence	Definition	Type(s) of evidence
Criterion-related	Test scores are systematically related to one or more outcome criteria. For example, can the assessment be used to draw inferences about a particular area on the basis of the results of the assessment?	Comparison of performance on the assessment against outside criteria such as • Grades • Class rank • Other tests • Teacher rankings
Content-related	Extent to which the assessment items represent the skills in the specified subject area.	Plan and procedures used in test construction For example: • Was a rational approach used that ensures appropriate content? • Did the process ensure that items would represent appropriate skills?
Construct-related	Extent to which the test measures the 'right' psychological constructs. For example, traits such as intelligence, self-esteem, creativity.	Demonstrate that the assessment items measure a single construct. For example: • Statistical analysis such as inter-item correlations, factor analysis • Correlations with measurements of related or different constructs

Other validity terms that are traditionally used in relation to research and the development of assessment tools are external, internal, and face validity. Internal and external validity relate to the appropriateness of a research design and confidence in the knowledge derived from it (French, Reynolds, & Swain, 2001). As such they are related to experimental research, although the terms are sometimes used in measurement research e.g. Lew et al. (2002). In this context, internal validity is interchangeable with content validity as it relates to the degree to which conclusions can be drawn, on the basis of the research design, about the causal effects of that one variable is having on another.

External validity, on the other hand, is the extent to which the results can be generalised beyond the research participants to explain what is occurring on other, similar situations (Schiavetti & Metz, 1997). In the case of assessment tool design, it is related to whether the tool can be used equally validly with other populations of subjects. Thus, external validity is

related to issues around sampling and sources of error when developing and using the assessment and is closely related to notions of reliability.

Face validity is sometimes confused with content validity (Schiavetti & Metz, 1997). A technical and a defensible judgement process is used to evaluate content validity whereas face validity is a subjective judgement made regarding whether the assessment appears valid. This judgement can affect both the assessor and assessee's engagement in the assessment process (Neary, 2000b) and is closely related to the concept of meaningfulness identified by Linn, Baker, & Dunbar (1991). These authors acknowledge the impact of the assessment upon learning as part of the validity considerations.

Thus the apparent meaningfulness of the assessment process may also affect how closely CEs follow the assessment guidelines and communicate with the student(s) about what they are expected to learn. Similar effects have been proposed in research on developing rating scales for rating employee performance where it was suggested that supervisors prefer to evaluate performance informally if the tool is not linked to important organisational and employee goals, leading to forms not being completed with rigour (Gomez-Mejia, 1988). This will affect how valid the test results will be and what interpretation and action should be taken.

3.3.1.2. Modern Validity Concepts

Messick (1989; 1994; 1996) developed an understanding of validity in testing that he argued is more unified and adequate than the traditional concepts identified in Table 3. The particular strengths of his operationalising of validity is that it takes into account the evidence of the value implications of score meaning as a foundation for action as well the social consequences of score use (Brualdi, 1999). This approach acknowledges that validating assessments is not only a scientific activity but has a political or social value as well. This social value arises as the evaluative judgements and decisions that are made through measurement will influence people and society (Messick, 1996).

Messick (1996) defined validity as follows:

> "Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *interpretations* and *actions* based on test scores or other modes of assessment." (pp. 1, author's italics)

Thus validity is not considered to be the property of an assessment but of the meaning of the test scores[4] that are generated through a combination of the assessment items, people doing the assessment, and the context of the assessment. The emphasis is on how the assessment scores are interpreted and used. The question for evaluating the validity of a testing or measurement instrument is: to what degree does score interpretation and use hold across all people or population groups and all settings or contexts? This is similar to McGuire's (1995) assertion that reliability and validity are the properties of particular interpretations of a performance on a specific assessment and not of a measurement technique or test format in itself. McGuire (1995) suggests that the only characteristics that can be assigned to a testing format are its intrinsic limitations and potential when it is used in an optimum fashion.

Messick (1996) considered the validity of an assessment to be an evolving property and that validation is a continuing process. He also cautioned that validity evidence should be considered a 'network' and will always be incomplete. Validation can only consist of making the most reasonable case to guide the current use of the assessment tool and future research to develop an understanding of what the assessment score may mean.

Messick (1996) outlined six distinguishable aspects of construct validity that are interdependent and complementary forms of validity evidence. These include:

1. Content Validity: Evidence that supports content relevance; representativeness; and technical quality of the assessment.
2. Substantive Validity: Theoretical rationales for observed consistencies in test responses; including process models of task performance and empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.
3. Structural Validity: Fidelity of scoring structure to the structure of the construct domain being assessed.
4. Generalisability: Extent to which score properties and interpretations generalise to and across population groups; settings; and task. Includes validity generalisation of test-criterion relationships.
5. External Validity: Convergent and discriminant evidence from multitrait-multimethod comparisons; and evidence of criterion relevance and applied utility.

[4] The term score is used to indicate any coding or summary of observed consistencies or performance regularities on any kind of assessment format. For example a test, questionnaire, observation procedure, work sample, portfolio, realistic problem simulation and so on (Messick, 1996).

6. Consequences: Assesses the value implications of score interpretation as a basis of action; as well as the actual and potential consequences of test use. Especially in regard to sources of invalidity related to issues of bias; fairness; and distributive justice.

Brualdi (1999) states that these six aspects of validity apply to all educational and psychological measurement with interpretations of scores, and the arising actions, explicitly or implicitly acknowledging or assuming these properties. Messick's (1989; 1994; 1996) unified approach to validity addresses all the previously identified components of traditional validity concepts as well as identifying a more complete range of evidence to be integrated into an overall validity judgement to justify score inferences and the subsequently implied actions.

3.3.1.3. Alternative Conceptions of Validity

Messick's approach has been critiqued by writers such as Wilson (1998) who suggests that Messick's validity concepts should be expanded by distributing the concepts across 13 different categories, all aiming to identify evidence of validity error. However, his discussion is founded in an epistemology that exists somewhere on a continuum between a qualitative and a quantitative understanding of the world. Quantitative researchers and measurement practitioners believe that there is an objective world that can be known and measured, including human qualities, and that there is one accurate representation of what exists – a true score. The qualitative end of the continuum extends to an understanding of the world as being entirely subjective such that there may be multiple meanings and versions of the truth, and all meanings are relative and have no material reality (Wood & Kroger, 2000).

This researcher acknowledges that truth and reality can be considered to be subjectively mediated by individuals (Crotty, 1998) and will address some of these considerations when addressing the role of judgement in assessment. However, ultimately the focus of this research is to meet the demonstrated need, in the most valid way possible, to have publicly accountable measures of the competence of speech pathology students to practice, prior to them entering the profession. This is similar to Carter's (1985) pragmatic comment regarding the subjectivity of assessment and the need to press on regardless:

> 'This is particularly true of the judgement that a person is fit to practice his profession. It is sometimes suggested that subjects should be excluded from a curriculum on the

grounds that they cannot be assessed. The realities of education for the professions do not allow the issue to be evaded in this way." (pp. 145)

In this respect, Messick's approach is the most practical description of the considerations that should be made when evaluating the validity of an assessment tool regardless of the inherent subjectivity of the judgements involved.

3.3.1.4. Summary

The process of determining the validity of an assessment involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations (AERA, 1999b). It is these interpretations that are evaluated not the test itself. Thus validation involves accumulating evidence to justify a statement of the proposed interpretation of test scores and rationale for how this interpretation relates to the intended use of the assessment tool. This network of evidence includes identifying the construct the assessment is intended to measure, the scope and extent of the construct, and a description of the conceptual framework for the assessment i.e. how the construct is distinguishable from other constructs and how it should relate to other variables (AERA, 1999b).

3.3.2. Reliability

Reliability of an assessment tool has traditionally been evaluated separately to validity, and this activity frequently attracts more time and effort than the evaluation of validity. This is probably due to the ease with which reliability issues can be mathematically represented and treated, and because it can be investigated on the basis of test data alone; as such, is easier to undertake (Feldt & Brennan, 1989). This tendency has been identified in the field of medical education and has been described as an 'obsession' that has occurred to the detriment or exclusion of assessment validity (McGuire, 1995). McGuire argues that it has resulted in assessment tools and formats that are reliable but of questionable validity in that it is not clear how performance on the assessment relates to actual functioning in the real world of health care delivery. Epstein & Hundert (2002) appear to be taking a similar stance when they identify that reliable assessments of core knowledge, abstract problem solving, and basic clinical skills have been developed for the medical field, but it is yet to be established that they also include the qualities of a good physician i.e. that they are valid.

There is no doubt that reliability is important; clearly we do wish to know whether speech pathology students can carry out their work equally competently at different times and in different settings. However, validity should have the highest priority as reliability is essentially integral to the validity of a test in that a valid test is reliable because it measures a real ability that is assumed to be stable (Friedman & Mennin, 1991).

Certainly Messick's view of validity subsumes reliability as it includes collecting evidence on the generalisability of assessment scores and can include information on the reliability of the assessment. Given that reliability of an assessment is an important consideration, albeit secondary to validity, and is frequently addressed in the assessment literature, it is useful to devote some consideration to its meaning.

Traditionally reliability analysis focuses on identifying and quantifying the inevitable errors that occur during measurement i.e. quantifying the inconsistency and consistency of examinee performance. It acknowledges that human performance is variable and that this will affect test scores due to a variety of factors, depending on the kind of measurement being used (Feldt & Brennan, 1989). Factors include variations in physical and mental efficiency of the test taker, uncontrollable fluctuations in external conditions under which the assessment is undertaken, instrumental variations such as equipment used and tasks required by the assessment that may favour one individual over another, and inconsistencies in the judgement of assessors.

The combined effect on a test taker's score is described as the error of measurement and can be positive or negative and systematic or random. The term 'measurement error' is strictly related to random error that cannot be controlled. Systematic errors consistently influence test scores positively and negatively and contribute to construct irrelevant variance but are not always controllable e.g. test taker anxiety (AERA, 1999a). Some sources of systematic errors can be identified and controlled such as eliminating items that favour one group of assessees over another e.g. according to gender, in a manner that is not relevant to the construct under examination.

A source of error that has been identified as a specific issue for assessment of students' performance in workplaces is the degree to which both CEs and students engage in the assessment. Cross and colleagues suggest that CEs are unwilling to fully engage with assessment tools or processes that they perceive as irrelevant or too unwieldy (Cross et al., 2001). This lack of engagement would impact both on the learning associated with assessment (Boud, 2000) and the validity and reliability of the assessment tool. Indeed, Neary (2000b)

found that both CEs and nursing students would choose to not use an assessment tool as intended if it was perceived as irrelevant to the practicum experience or jargonistic which could affect the reliability and validity of the assessment process.

Error of measurement is based on the notion of 'true score' such that the error is the hypothetical difference between the assessees' observed score and their true score for the procedure (AERA, 1999a). The 'true score' (or 'universe score' in generalisability theory) is conceptualised as a personal parameter that remains constant over the time it takes to take several measurements (Feldt & Brennan, 1989). Investigating reliability, or error of measurement, includes gathering and analysing data to identify and quantify major sources of error (AERA, 1999a) and the degree of generalisability of scores across alternate forms, scorers etc. In addition, it is recommended that a description of the population be provided, as the assessment data may relate specifically to that population (AERA, 1999a).

In summary, traditional notions of reliability and error of measurement are important components of the overall validity[5] of an assessment to the extent that they contribute to an understanding of a justifiable interpretation of test scores. A case was made earlier in this discussion in favour of continuing with the current practice of speech pathology preparation programs in valuing the evidence provided by workplace based performance assessments, regarding their students' readiness to enter their profession. It would appear that an exaggerated focus on reliability, and concerns regarding controlling error, can lead to a preference for assessments of competence (e.g. OSCEs) rather than evaluating performance in the workplace; this is to the detriment of validity. However, much of the concern regarding using assessments of performance, rather than competence, is due to concerns that performance assessments have particular drawbacks with regard to evaluating their validity.

3.3.3. Performance Assessments

3.3.3.1. Defining Performance Assessment

The term 'performance assessment' and its related validity issues covers a range of assessment types, many of which are more closely related to assessing components of underlying competence as a predictor of workplace competence, rather than performance in the sense of assessing performance in real workplace action. For example, performance

[5] The term 'validity' will now be used in this thesis on the assumption that includes traditional concepts of reliability.

assessment is described by Gronlund (2003) as being constructed to different degrees of 'realism' and includes assessment tasks that mimic the real life task rather than actual workplace performance.

Generally it appears that, in the educational assessment literature, a continuum of assessment types are conceptualised, according to the degree of realism of the task and integration of knowledge and skills required to complete it successfully. Performance assessments are assessment tasks that require integration of knowledge and skills as well as the students' maturing grasp of underlying concepts (Masters et al., 1999) as opposed to simplistic sampling of specific, decontextualised academic knowledge. They aim to provide assessment tasks that evaluate the quality of the processes students engage in, or the product they produce or a combination; and are typically at the less structured end of the response continuum (Messick, 1996).

In this sense a pencil and paper test can still be classified as a performance test if it involves a complex, realistic task (Gronlund, 2003). This viewpoint is echoed by Schuwirth & van der Vleuten (2003) and Page (2004) who identify that even multiple choice questions can be constructed in a way that provide realistic tasks and sample complex integration of knowledge required for effective medical practice. However it is clear that, while the format of an assessment does not determine what is assessed, some formats are more suited to assessing particular types of knowledge than others (Schuwirth & van der Vleuten, 2003).

The case has already been made for the need to assess performance in the workplace when deciding if speech pathology students are sufficiently competent to enter the profession in addition to more decontextualised assessments within the educational program. Messick (1996) suggests that workplace assessments would fall into the subset of 'authentic' assessments but this distinction contributes little to increasing clarity as all performance assessments (in the broader sense of the term) appear to be an attempt to be more authentic. However, he does make a useful distinction between construct versus task driven performance assessment.

Task centred assessment either sees the performance of a specific task as the focus of the attention, or as specific demonstration of more general constructs of interest e.g. problem solving, or other skills that can be generalised to the specific task. Messick (1996) cautions that task centred assessments can result in scoring criteria being tailored, sometimes in an unacknowledged fashion, to the properties of the task. This will limit the generalisability of the assessment information as the constructs are elicited in a task dependent manner. This is

in fact reminiscent of prior discussion regarding critiques of CBE and assessment and the need to define competence in generic as well as occupational terms. Messick recommends taking a construct focussed approach to the design and scoring of performance assessments for reasons of validity expounded on further below.

3.3.3.2. Validity and Performance assessments

As described earlier, the belief that it is not possible to adequately ensure the validity of workplace based performance assessments has led to the use of more standardised, controlled assessment types to simulate work skills and this has paradoxically been to the detriment of validity, in particular some aspects of generalisability. This has been no more evident than the dilemma regarding the fact that it cannot be assumed that complex skills can be generalised from one context to another (Friedman & Mennin, 1991) while at the same time it is acknowledged that traditional reliabilities may be low due to the variability of raters (Cross et al., 2001). Linn et al. (1991) argued that, as a result, performance assessments require an expanded framework to evaluate validity and should include Messick's focus on consequences of assessment but also notions of fairness, transfer and generalisability, cognitive complexity, content quality and coverage, meaningfulness, and justifications of their cost.

However, Messick (1994; 1996) counters that these criteria do not provide more information on the validity of an assessment than his suggested criteria. He argues that the two major threats to validity of all assessments are construct irrelevant variance and construct under representation, in other words, assessments being too narrow or too broad. Thus the primary validation concern is the extent to which the assessment may under represent the construct of interest while at the same time introducing measurement error through construct irrelevant variance. Messick puts a detailed case as to how these two threats to validity can be carefully evaluated through application of the six interrelated aspects of validity to performance assessments. This will form the basis for future discussion regarding considerations for assessment design.

It would appear that Messick's argument in support of the application of his validity criteria to performance assessments has subsequently gained acceptance by authors such as Linn and others (Miller & Linn, 2000). However, performance assessments do have issues particular to them, for example, possible advantages due to better construct representation

(Messick, 1996) but particular challenges with regard to validity criteria such as generalisability (Miller & Linn, 2000).

3.3.4. The Role of Judgement in Assessment

The role of judgement in performance assessments is seen as contributing significantly to the difficulties in generalising the implications of score performance and therefore having significant impact upon the validity of an assessment tool. This has probably contributed to a strong focus on investigating the reliability of tools assessing performance and, given the high stakes nature of most assessments, this is not unreasonable. In particular, concern is expressed over the role of judgement in determining whether students 'performances are at a particular level and the subjective influences upon this judgement (Alexander, 1996; Chapman, 1998). This concern is echoed by CEs who express concerns that their judgement has been influenced by irrelevant personality factors (Duke, 1996) and reviews of the literature on rating performance which suggest that there are multiple sources of error involved (Landy & Farr, 1980).

This concern regarding the subjective judgemental nature of assessment unduly affecting the evaluation of students' performance is not well supported in the literature. While Alexander (1996) found evidence that assessors make subjective judgements about students and that these judgements influence assessment grades in combination with consideration of assessment criteria and can sometimes be erroneous, other research that has investigated sources of error has not found evidence in support of this position. Friedman & Mennin suggested as early as 1991 that the judgement of raters was a smaller source of error than other factors such as case specificity – the well documented effect of student performance varying considerably from case to case on an OSCE style examination. This point of view has been supported by generalisability studies that use a matrix to examine the influence of various factors upon the scores received by the students. These studies have found that the rater or judges' behaviour generally had a much smaller effect than other factors such as assessee knowledge and tasks sampled (Govaerts et al., 2002; Keen, Klein, & Alexander, 2003; Ramsey et al., 1993; Shavelson, Gao, & Baxter, 1993).

This is perhaps not as surprising as it would appear if one accepts the viewpoint of writers such as Hager (2000) who argues that the development of professional judgement is integral to the growth of expertise within the workplace. Hager (1999) defines judgement as deciding what to believe or do after taking into account a variety of relevant factors and then acting

accordingly; a process that is identified in the professional practice of speech pathology (SPAA, 2001). It would therefore be expected that the raters involved in such studies would be able to make appropriate judgements of assessee performance given that the assessors are at least practitioners, if not acknowledged experts, in the field of professional activity being judged.

It is also relevant to note that the idea that assessment of any human being can be totally scientific and objective has been described as a myth (Bitzer, 1999). Leach, Neutze & Zepke (2001) point out that the all assessment is inherently subjective and cite research in support of this view that identifies wide variation in grading of essays can occur between judges. Jones (2001b) found that even if judgements are based on highly specified competency based criteria the process was an ethical and attentive one. Jones highlighted that even expert and highly trained judges, such as the Australian Supreme Court, are rarely unanimous in their decisions.

A similar viewpoint is espoused by Moss (1994) who describes this as a difference between hermeneutic and psychometric approaches to assessment. She describes the hermeneutic approach as involving a holistic and integrative interpretation of human phenomena that attempts to understand the whole in the light of its parts. The process includes testing this interpretation against available evidence until all aspects of the evidence can be coherently interpreted as parts of the whole performance. In contrast, the psychometric approach interprets the individual's performance through aggregating the scores derived from judgements of decontextualised samples of behaviours that are not related to or interpreted in the light of other relevant contextual factors. Moss therefore suggests that judgement is an important part of the process of assessment and safeguards the validity of the process. She also points out that many higher education high stakes assessments rely strongly on judgement and do not concern themselves overly with the reliability of judges' assessments e.g. awarding of post graduate degrees (Moss, 1994).

In fact, it can be argued that judgement is both inevitable and integral to the design of any valid assessment tools e.g. content is often determined by panels of expert judges (AERA, 1999b; Messick, 1989). As such, performance assessments would not appear to be any more disadvantaged by the need for judgement than any other assessment technique. Ultimately, as for any assessment approach, issues related to generalisability and fairness of performance assessments must be attended to during assessment design (Moss, 1994) and the related sources of error controlled for and evaluated, or indeed, judged.

3.4. Summary

This research endeavours to shed some light on the assessment of competence, a task summarised somewhat dauntingly by Wass (2001) as follows:

> "Assessment at the apex of Miller's pyramid, the "does", is the international challenge of the century for all involved in clinical competence-testing. The development of reliable measurements of student performance with predictive validity of subsequent clinical competencies and a simultaneous educational role is a gold standard yet to be achieved." (pp. 948)

This literature review has drawn together information regarding the nature of professional competence and assessment that are important for informing the development of a tool to assess whether speech pathology students are ready to enter the profession. First, competent performance of the speech pathology profession requires appropriate professional action both in the present, and into the future. It relies on the competent exercise of complex professional judgement and action resulting from integrated combinations of knowledge, skills, and personal qualities. This professional judgement must be exercised across all tasks and contexts of the profession and includes performance of occupational competencies as well as engaging in behaviours indicative of generic professional competencies. These generic competencies are considered to be instrumental in the development and maintenance of occupational competencies over the speech pathologist's working life.

Second, professional action is complex and needs to be applied to different workplaces. It will change and develop over time and fluctuate in quality due to a range of factors including personal and workplace characteristics. Thus assessment to determine whether professional competencies are both present now and likely to be maintained in the future is, in itself, a matter of a considered exercise of professional judgement on the part of the assessor. Ensuring that this assessment is valid is a matter of careful attention to the content, format, and process of assessment such that the assessment tool effectively informs and supports the process of gathering and weighing evidence to enable a judgement that is both valid and justifiable in terms of its consequences. These consequences include both the impact upon the students' learning as well as the high stakes decision regarding whether students are sufficiently competent to enter the profession.

Finally, workplace assessment of competency, defined as the inferring of competency through the observation and judgement of the speech pathology students' behaviours when carrying out tasks in the actual real world arena of professional practice, is recommended. Other methods rely on unproven links between decontextualised assessment tasks and appropriate workplace performance under representing the complex construct of professional competency. Adequate construct representativeness is critical to ensuring appropriate assessment consequences related to learning as well as the final decision regarding the students' eligibility for entry into the speech pathology profession. Workplace assessment introduces disadvantages that threaten validity as well; these aspects require careful consideration during the design process and acknowledgement when interpreting assessment results. However such threats to validity are not avoided by carrying out assessments in other contexts or forms.

These conclusions led to the decision to develop the assessment tool as a workplace based assessment to support the judgement of competency by CEs. In the Australian context CEs have the opportunity to observe a student(s) on multiple occasions over time and are responsible for facilitating the student(s)' learning and contributing to the assessment of his/her competence. The research project was conducted in two phases and information regarding ethics approvals for the entire research process can be found in Appendices 1, 2, 3 and 4. The first phase focussed on the design of the assessment content and process to maximise its validity. The second phase involved field testing the validity of the assessment.

PHASE 1: DESIGNING THE ASSESSMENT

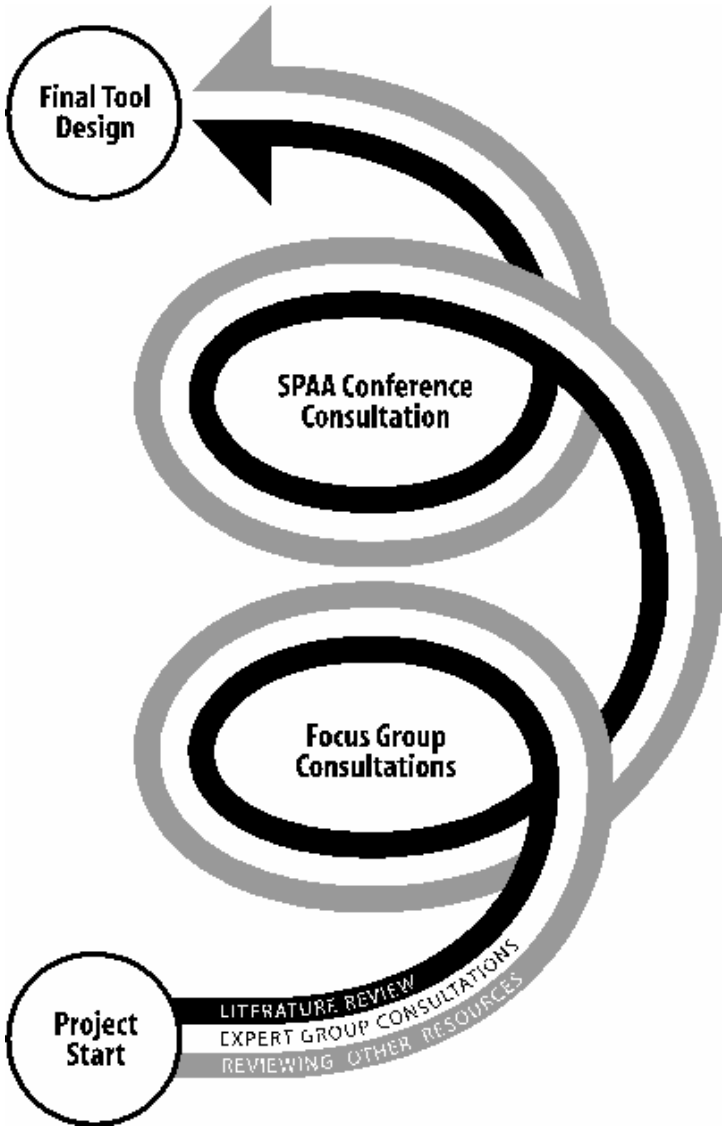
CHAPTER FOUR

4. CONSULTATION REGARDING ASSESSMENT DESIGN

4.1. Overview

Assessment design was a reiterative process involving multiple consultations with various experts and stakeholders, informed at all times by ongoing analysis of the literature on competencies and their assessment, current assessment tools, and resources such as the CBOS (SPAA, 2001). Fig. 1 is a diagrammatic representation of this process with each loop of the spiral representing a formal consultation activity. Each consultation was informed by concurrent review and analysis of the literature and other resources.

Figure 1. Tool design process



Each component of the consultative process will be described in terms of its method and results (Chapter 4). Chapter 5 will then summarise the assessment design considerations, integrate the consultation results with information derived from the literature review and other resources, and describe the assessment tool.

4.2. Consultations

4.2.1. Expert Group

Using expert groups to contribute to the development of assessment tools is a well established tradition in assessment design e.g. AERA (1999a); Clauser (2000); and Miller & Linn (2000). Expert opinion is sought to promote the validity of the instrument through the development and critical examination of the theoretical framework proposed for the construct to be assessed, guiding the development of assessment content, assessing how well the assessment content and processes represent the construct of interest, and to critically examine scoring rules and criteria. A range of experts were consulted throughout the development of the assessment format, one group in particular provided ongoing guidance.

The funding proposal to support this research was put forward by four experts in the field of clinical education and speech pathology. It was this group of experts that assisted with the development of the overall research process, but also provided specific input and feedback as to the assessment design process. The four people were:

1. Michelle Lincoln (Principal supervisor), PhD, Senior Lecturer and Director of Clinical Education, School of Communication Sciences and Disorders, The University of Sydney.
2. Alison Ferguson (Co supervisor}, PhD, Associate Professor Speech Pathology Program Coordinator, The University of Newcastle.
3. Lindy McAllister (Co supervisor), PhD, Associate Professor, Speech Pathology, Charles Sturt University.
4. Paul Hagler, PhD, Professor of Speech Pathology and Audiology, The University of Alberta.

Dr Hagler was able to provide input via email and during face-to-face discussions of the project with Dr McAllister. Drs Lincoln, Ferguson and McAllister provided expert opinion via telephone, email, and face-to-face consultation meetings throughout the research process. These meetings finalised consensus positions regarding specific issues related to research

procedure and tool design and identified issues that required further consideration. Dr Lincoln, as principal supervisor, met with the researcher by telephone every 2 to 4 weeks and communicated via email between meetings.

4.2.2. Focus Groups

4.2.2.1. Background

A decision was made soon after commencement of the research project to seek information from students and university and field CEs to inform assessment design in the first instance. This decision was taken for three reasons. First, broadening the consultation beyond the expert group right from the start of the project was considered important given that researchers such as Cross (1998) have identified that CEs may have a different perception to university staff as to what are key competencies for professional practice. As identified in the literature review, ensuring that the content of the assessment was meaningful to students and CEs is also critical to ensuring their meaningful engagement in the assessment process – an important influence on validity. This is closely aligned with safeguarding the values of assessment fairness (Lew et al., 2002) and empowerment of students within the assessment process (Leach et al., 2001) and considered to be important to ensure that both intended and unintended outcomes of the assessment process were identified and addressed from the outset.

Second, information gained from initial literature reviews highlighted that, while the CBOS with its comprehensive outline of occupationally specific occupational competencies clearly needed to be integral to the assessment of competency, there may be other generic skills that enabled the expression, flexible application, and lifelong development of these occupational competencies. Consulting with CEs and students via focus groups was proposed as one strategy to assist in developing an understanding of what these generic competencies might be and how to include them in the assessment. Finally, if assessment is a matter of considered professional judgement on the part of the assessor, then it was important to consult with potential users of the assessment format as to how the content, format, and process of the assessment could be designed to effectively inform and support their judgement. This was also important to ensure that the assessment would be practical in the workplace.

4.2.2.2. Method

Focus groups were selected as the method for seeking input from the three stakeholder groups (students, CEs based in the field and at universities) identified as possibly holding

divergent but important and relevant views regarding the optimum assessment content, format, and process. Focus groups are generally defined as a series of interviews with groups of people who possess similar characteristics and who provide qualitative data in a focussed discussion (de Laine, 1997). The advantage of focus groups over interviewing individuals lies in the social effect of groups. The group context means that the participants are able to hear each other's ideas and use these in formulating and elaborating their own opinions and values and can encourage a greater variety of communication than other methods of data collection. This variety and process of elaboration can actively contribute to clarification of ideas and operational aspects of a research process (de Laine, 1997).

Focus group participants.

Focus groups were organised with a view to representing students from multiple university programs; country and metropolitan CEs working with a variety of client groups and service delivery models; as well as academics and CEs from a range of university programs. CEs in country centres generally had worked with students from a number of different university programs. Given the distances and travel costs involved, focus groups were conducted by the researcher via teleconferences and as well as face to face and also in person by members of the expert group. Appendix 5 provides information on the groups interviewed.

Krueger & Morgan (1998) suggest that an ideal focus group size is around 6 to 10 participants but that smaller groups are effective if there is a high level of involvement in or knowledge about the topic, or the topic is complex – which was true for these focus groups. While it was aimed to have at least 5 participants this did not always occur; however the amount of discussion generated and information gained did not relate to the number of participants. It was difficult to recruit sufficient student numbers to form a focus group so semi-structured interviews were conducted using the same question format (described in the next section) with 2 pairs of students, and another single student.

Generally focus group research involves a series of 4 or 5 groups (de Laine, 1997) or repeated until no new information is gained (Krueger & Morgan, 1998). Analysis of the data yielded by the 6 focus groups of CEs and 3 semi structured interviews with students revealed strong similarities in information gained from all groups represented (regardless of size and method of interview) and the expert reference group. This resulted in strong confidence that the process had allowed for effective and inclusive data gathering and no further focus groups or semi structured interviews were required.

Process.

Question Development.

The content of the questions was guided by the overall aim of canvassing as wide as possible range of ideas and options for the assessment tool including consideration of assessment practices other than those currently used by Australian programs. Indeed, it was the students' and CEs' views on the issue that were sought not their response to the researcher's views (Stewart & Shamdasani, 1990). Stewart & Shamdasani's (1990) guidelines for developing the interview questions were used and included ensuring minimal structure, not suggesting potential responses, ordering questions from the more general to the more specific and relative to their importance to the research agenda, allowing for flexibility in pursuing lines of enquiry, anticipating lines of discussion, and aiming for clear wording.

Stewart & Shamdasani (1990) also advise trialling the interview guide prior to using it. Moderators were asked to feedback any issues to the researcher regarding the question guide after using it with a focus group with a view to modifying the process or questions if required. This was not found to be necessary.

The researcher aimed to explore possibilities regarding tool format or design and content, based on issues raised in the literature review regarding the nature of valid performance assessment, generic and occupational competencies, and the role of judgment in assessment. Specifically these included developing an understanding of CEs' and students' perspectives on:

1. What should be included in the assessment.
2. What formats could be considered.
3. What, if any, generic competencies or other dimensions of performance should be assessed in addition to the CBOS occupational competencies.
4. What content and processes were perceived as contributing to validity or fairness, including supporting judgment.
5. How competence develops.

A question guide was developed for moderators that consisted of three major questions and suggested follow-up questions, wording was slightly modified for student and CE groups to reflect their different perspectives on the assessment process (Appendix 6).

Conducting focus groups

Moderators were provided with suggested procedures for the practical aspects of running the focus groups as well as ensuring that information and demographic sheets were distributed and read (Appendices 7, 8 and 9). As per ethics approval requirements, an information sheet was provided to participants describing the research, how the data from the focus groups would be used, assurances of confidentiality, and contact numbers if the participants wished to express concerns regarding the conduct of the groups or the research (Appendices 8 and 9). Focus group members were advised that their participation in the discussions implied their consent for this information to be used in the research.

Recording data

It was intended that each group/discussion be tape recorded for the researcher to transcribe. Unfortunately equipment failures (despite prior testing) resulted in data from two groups being generated from field notes only (one pair of students and field educators from Canberra) and data from the University of Newcastle CEs was a combination of field notes and transcription of a poor quality recording. The remaining data was transcribed from tape recordings of the group discussions.

Analysis

A thematic analysis of the transcripts and field notes was conducted. Each statement was summarised by a key word or phrase that was then collated into a summary of the key concepts and issues identified by participants. The source of each statement was identified so it could be determined whether the issue was held in common or specific to students, field or university clinical educators. This summary was then examined and themed categories identified that accounted for all the concepts and issues. A similar process was undertaken independently by Dr Michelle Lincoln, and any differences in categorisation were identified and resolved. The final categorisation of the transcripts was also examined by Dr Alison Ferguson and Dr Lindy McAllister.

4.2.2.3. Results

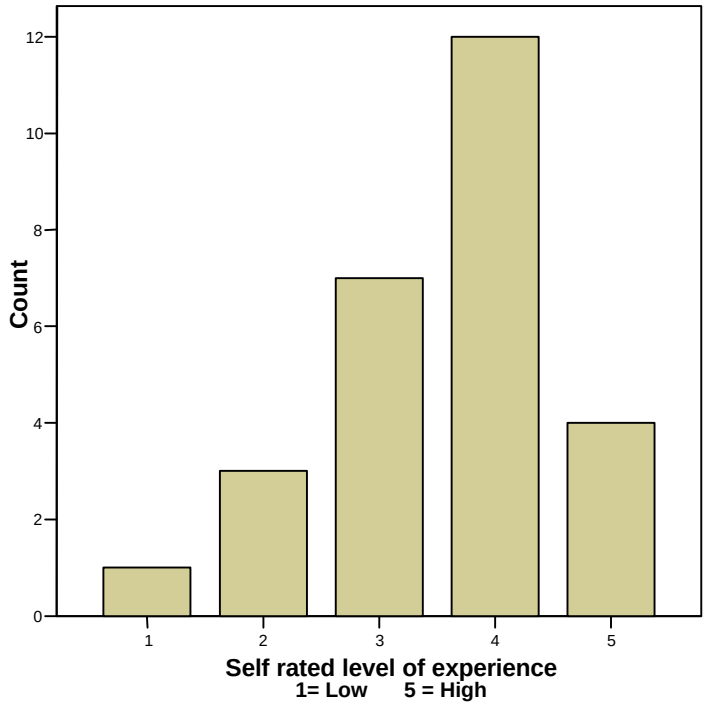
Demographics.

Twenty-nine of the 31 CEs who participated, returned demographic data on the nature of their experience as CEs and experience as speech pathologists. Their experience post graduation varied from 3 to 31 years, with 11 having worked more than 10 years as speech pathologists. Around half (15) had supervised students from one university only, with the

remainder (14) having supervised students from two or more speech pathology programs. The majority (25) had worked with students from more than one year level and considered themselves to be very experienced (Fig. 2) and the same number had received training in clinical education or assessment.

Four students were entering their fourth year of a 4 year degree program in speech pathology and had had experience with at least two assessments of their workplace performance. The fifth student was entering her third year of the course and had been assessed on one placement.

Figure 2. Self-rated level of experience of clinical educators participating in focus groups



Results

Data gathered during the focus group phase of the research contributed significantly to the reiterative process of tool development. It both confirmed the initial thoughts of the researcher and the expert group, e.g. how to support judgment, and elaborated further on issues such as what competencies to include in the assessment. The focus group data therefore was a strong influence on the researcher's development of a theoretical framework for the assessment and guided ongoing literature review, expert group discussion as well as the first draft of the tool design presented for further discussion at the SPAA 2002 conference (described subsequently). This information was synthesised with other sources to inform tool design as described in the following chapter (Chapter Five).

In summary, there was a high degree of agreement between all researchers on the themes represented by the focus group/semi structured interview data.The groups showed strong convergence in their opinions regarding assessment design regardless of whether they were students, field, or university CEs. All groups made the following points:

1. Both occupational (CBOS) and generic competencies were relevant to assessment of competency.
2. Clear, detailed criteria and examples to guide ratings were required. All felt that training and peer review was important to ensure consistency of ratings across placements.
3. Clear definitions of what is being rated are essential.
4. Must be applicable to different caseload and placement complexity and student experience.
5. Student performance should be rated on a range of features e.g. not on independence only.
6. The rating scale should better reflect progress over time, within and across placements, than do current scales used in assessments.
7. Ultimately the assessment process is subjective.
8. The assessment impacts on learning and this needs to be attended to.
9. Should include a formative and summative component with students being involved in the assessment process.
10. Opportunity to make or receive comments was highly valued.

All these issues, with the exception of the specific issue of being able to include comments in the assessment, had been previously identified by the researcher and discussed by the expert group. This confirmed that their understanding of the research task was congruent with that of their colleagues and students. A great deal of information was collected on what might constitute generic skills in speech pathology and groups tended to nominate similar ones e.g. interpersonal communication, lifelong learning skills, critical thinking. In addition, a number of CE groups reflected on the tension between detail and the need for brevity and suggested optional layers of details and resources, confirming initial discussions regarding tool design by the expert group. CEs tended to devote more time to discussing the teaching aspect of assessment but students acknowledged it as an issue as well. Some students and CEs also raised the suggestion of a computer-based version of an assessment tool. CEs in the field and students were both concerned regarding expectations for different placements and clients but this issue was not raised by university based CEs.

In summary, the focus groups' opinions showed a great deal of convergence with each other and the research group and were a rich source of data to inform tool design, as will be discussed later in Chapter Five.

4.2.3. Speech Pathology Australia Conference Consultation

4.2.3.1. Background

Speech Pathology Association of Australia Ltd. (SPAA) convenes an annual conference each May. The 2002 conference in Alice Springs was identified as an opportunity to involve a broader range of speech pathologists with an interest in clinical education in the design of the tool content, format and process. The conferences are organised in a module format where convenors volunteer to coordinate a set of papers on a particular theme. Proposals for combined papers and consultation sessions were submitted and accepted to two modules: "Clinical Education in the 21st Century: Challenges and changes" and "From novice to expert: Creating and maintaining lifelong learners". The papers/forums submitted aimed to both provide CEs with information regarding reliable and valid assessment and the promotion of life long learning. An additional aim was to seek feedback on the draft assessment tool design and further input on the nature of the generic competencies thought to be a part of quality speech pathology practice.

4.2.3.2. Method

Procedure for presentations

The researcher presented two papers/discussion forums, each structured as a short paper followed by small group discussions moderated by the researcher as well as members of the expert reference group, and on one occasion, Dr Alison Russell. Dr Russell is course coordinator at Flinders University and teaches and practices in clinical education. Moderators were provided with a briefing by the researcher that outlined their role, the guidelines for the task, and timelines for the session.

As per ethics requirements, session participants were provided with the research information sheet explaining the research, the aims of the consultation, and their rights, as well as who to contact if concerned about any aspect of the research or how it was conducted (Appendix 10). They were informed that their consent to use data from the discussions was implied if they participated in discussions. A demographic sheet and contact details sheet were also distributed for participants to fill out and return if they wished to receive transcripts

of the discussions or further information on the project (Appendix 10*)*. The questions for discussion were developed by the researcher in consultation with the expert group and were structured to stimulate focused discussion and generation of information. Moderators recorded the discussion with field notes and scribes, who were selected and briefed prior to the forum, summarised discussions on large sheets of paper displayed in front of each group while discussion was occurring.

The first presentation and discussion took an hour as follows:

1. Brief introduction to session and proposed model for the assessment process and invitation to comment on the draft format (10 minutes).
2. Break into 4 groups for reflection and discussion of questions related to the proposed assessment format as per Table 4 (total 40 minutes).
3. Larger group discussion/reflection facilitated by Sue McAllister (10 minutes).

Table 4. Discussion Questions and Tasks for "Making assessment reliable, valid and achievable-Forum"

Primary question/task	Sub questions/tasks
What should be the single dimension to rate the CBOS elements on?	• Describe/define the dimension • What aspects of performance does it include? • Name the dimension
What are the underpinning skills that should be rated /commented on?	• Brainstorm all the possible underpinning skills* that should be considered. • Are any of these clustered or linked? • How many of these would you consider have to be included in an assessment of a student's competence? • Rank the top 3 underpinning skills.
Discussion/Reflection	• Comments on the key underpinning skills to be considered. • What weighting should these have as compared to more specific CBOS based skills? • Do we feel we can judge on these dimensions? • Are the underpinning skills too value laden to make it feasible to judge? • How well does the suggested tool design lend itself to being rated on these dimensions? • Does it meet the criteria of 'do-ability' and supporting reliable judgements of competence? • Other comments on the tool design

* These were defined as and were later termed 'generic competencies'

The second presentation and discussion took approximately 40 minutes as follows:

1. Brief introduction to session and proposed assessment format with a focus on the aspects related to lifelong learning skills (10 minutes).
2. Break into four groups for reflection and discussion of questions related to the definition and assessment of lifelong learning skills as per Table 5 (30 minutes).
3. Close (1 minute).

Table 5. Discussion Questions/Tasks for "Creating life-long learners: How can we be sure we have?"

Primary question/task	Sub questions/tasks
What is the package of skills/knowledge/attitudes or other attributes a new graduate needs to bring to the workplace to ensure lifelong learning?	• Brainstorm a list • Identify the key or indispensable features i.e. what might be most important to the workplace/manager? • How do you know if the person is a lifelong learner?
Reflect on the previous task through discussion of the following questions	• What behaviours can be observed that would demonstrate these attributes? • What would be the indicators that these are emerging? • When do we decide someone can be 'safely' described as a lifelong learner? • Are lifelong learners born or made? • What strategies can a clinical educator/manager use to assist a student/employee to develop lifelong learner skills/knowledge/attitudes?

The field notes from moderators and scribes for each group were summarised and circulated to those who indicated that they wished to receive them, and comment invited.

4.2.3.3. Results

Demographics

Not all participants returned demographic data and, as discussion moderators were not asked to count the number of participants in their discussion group, exact numbers of participants can not be determined. It was estimated that approximately 25 to 30 people were at each session. Demographic data was returned by 24 participants at the lifelong learning session (LL session) and 21 from the clinical education session (CE session).

The groups varied slightly, which was not unexpected given one session was primarily focused on clinical education however the lifelong learning presentations covered post graduate perspectives as well and so were of interest to managers as well as educators. As can

be seen by Figs. 3 and 4, the clinical education group was more likely to have supervised large numbers of students and to rate themselves as experienced. In addition 91% had attended training in clinical education compared to 68% of the LL session, 68% had post graduate qualifications compared to 54%, and 62% were or had been employed as a CE versus 42% of the LL session. However, overall both groups represented a wide range of experience. Participants had worked from 1 to 33 years across a wide range of workplaces and the overwhelming majority had supervised students from more than one year level with around 40% of both groups having worked with students from more than one university program. Both groups had quite a number of speech pathologists graduated from programs in countries other than Australia: 21% (6) participating in the LL session; and 29% (5) in the CE session. Five of the 24 participants in the LL session and 3 of the 21 in the CE session supervised students from non-Australian programs.

Figure 3. SPAA Conference consultation groups by number of students participants have supervised

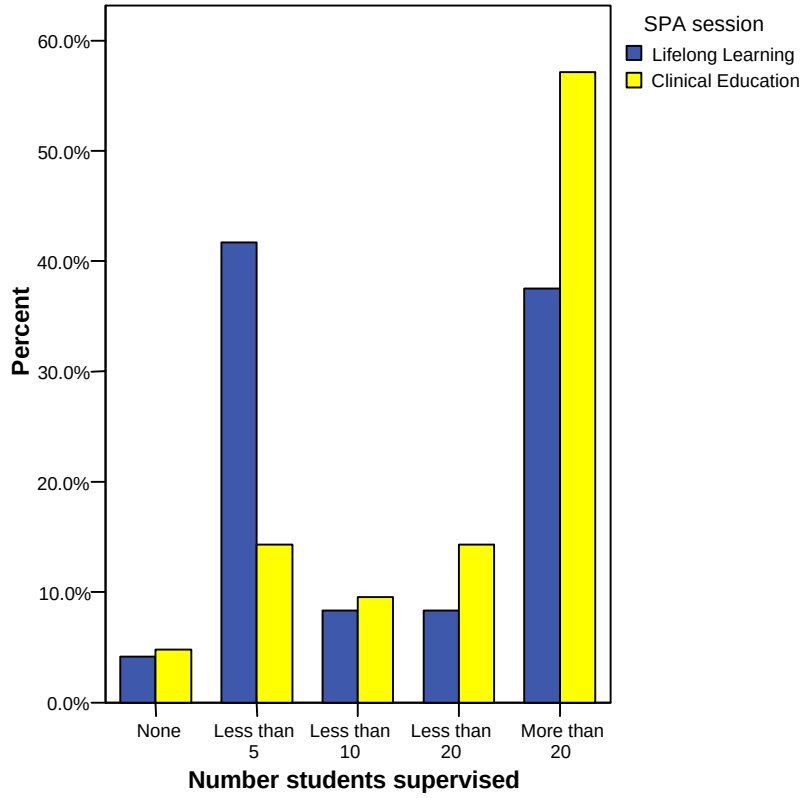
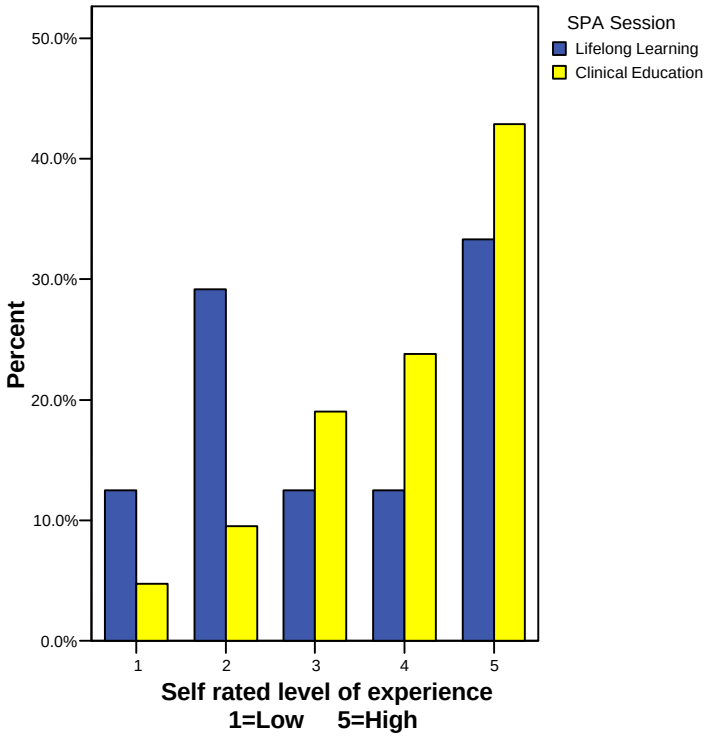


Figure 4. SPAA consultation groups by self rated expertise



Results

The SPAA Conference consultation, similarly to the focus group data, yielded a great deal of information that contributed significantly to the reiterative process of assessment development. The data both confirmed and elaborated information received from the focus groups, the literature, and expert group consultations and was integrated with other sources to guide assessment design, as described in Chapter Five. Those participants who wished to receive a summary of the discussions were sent one via email or hard copy and invited to contact the researcher if the summaries did not accurately reflect the discussions. No feedback was received regarding the content. The following is a summary of the information gathered.

Making assessment reliable, valid and achievable – Forum

No critique was offered of the proposed assessment model at the forum or subsequently with only positive comments received, affirming the suggested direction for the tool design. Conference participants divided into four groups of around 6 to 7 participants in each. Three out of the four groups addressed discussion to the first two tasks in Table 4 with the fourth group briefly covering the third task. It was clear that all participants felt that student performance needs to be assessed on more than one dimension. Although one group did not generate much discussion around this first question except to identify the degree of guidance

students require as being a useful dimension to consider and to consider how the scale might reflect increasing competence. It was apparent that the discussion tended to move from seeking a dimension or continuum on which to rate student performance on the CBOS skills towards identifying generic competencies of relevance to assessing performance. This may have been due to the difficulty of this task or to confusion engendered by the use of the term 'dimension' rather than something more explicit such as 'continuum'. The dimensions that were identified by the three groups included the rating performance with regard to:

1. Independence and/or interdependence: These reflected ideas about the need for students to demonstrate developing and appropriate autonomy in their work practice.
2. Adaptability or flexibility.
3. Ability to synthesise or integrate aspects of performance including being able to develop a strategic overview or wider perspective and moving from isolated practice of simple tasks to integrated practise of complex tasks.

In addition, two aspects of performance that perhaps could be better described as generic competencies than features to describe increasing competency on a task were discussed:

1. Lifelong learning: a number of concepts clustered around ideas of lifelong learning such as the rating including the students learning abilities.
2. Development of critical or creative thinking, or clinical reasoning.

Concepts of efficiency were mentioned by two groups; however both identified that this was difficult to both define and quantify and perhaps could be subsumed in other dimensions. Two groups also mentioned that the ratings should include interpersonal skills; which also more usefully lends itself to a generic competency on which to be assessed.

The second question, "What are the underpinning skills that should be rated/commented on?", generated very detailed lists of skills considered to be relevant to the competent practice of speech pathology. Groups were asked to summarise these into the top 3 'must have' competencies and a remarkable degree of commonality existed between the groups, with four main competencies being mentioned (summarised in Table 6, see Appendix 11 for the full listing of these competencies by group).

Table 6. Consensus on Generic Competencies From SPAA Consultation

Generic Competency	Sub-competencies
Clinical Reasoning Skills	• Judgement/decision-making/problem solving • Moving from theory to practise • Ethical reasoning • Reflection (links to lifelong learning skills)
Lifelong learning skills	• Self-evaluation, reflection, change and learning
	• Reflective practitioner including: lifelong learning, self-evaluation, self praise
Professionalism	Handling contextual issues
	• Subsumes – organisational skills, ethics • Personal/self management
Interpersonal skills	Communication skills
	• Interaction/interpersonal skills

Two further skills were mentioned, by only one group, that did not fit under this categorisation: implementation skills and intercultural competence. However, implementation skills are covered in detail by the CBOS and it could be argued that intercultural competence could be subsumed under interpersonal and/or clinical reasoning skills.

Creating life-long learners: How can we be sure we have?

Conference participants again divided into four groups of around 6 to 7 participants in each. Discussion was very extensive, generating detailed lists of the attributes that identify a lifelong learner, how these attributes might be expressed in behaviours and be recognised as emerging. A great deal of commonality again existed between the groups' conceptualising of lifelong learning skills with the themes summarised in Table 7 existing within all discussions.

Table 7. Summary of Lifelong Learning Skills Identified at SPAA Conference Consultation

Lifelong Learning skill	Sub-skills
Motivation	• Desire to learn • Willingness to take risks • Initiates/takes responsibility for learning • Predisposition to learning/open-mindedness/curiosity • Planning for own learning
Self awareness	• Understanding own learning style • Knows limits to current knowledge • Able to separate personal from professional in relation to receiving feedback
Reflective skills.	• Self evaluation • During and after professional action
Ability to apply new learning	• Change performance • Adaptability/flexibility
Problem solving	• Identify and analyse the problem and possible solutions • Critical thinking • Planning to resolve a problem
Accessing and using resources effectively	• Asking questions appropriately • Researching • Listening
Synthesis	• Able to synthesise information from different sources including analysing or structuring information effectively • Able to see the big picture • Integrates new knowledge

The second part of the discussion focussed on how to assess if someone is a lifelong learner and was framed as identifying how you know if someone is a lifelong learner, what behaviours would demonstrate a lifelong learning orientation and indicate this emerging ability. Again discussion was both broad and detailed, with several behaviours identified by all groups as demonstrating a lifelong learning orientation. These included that students needed to understand their own learning needs and demonstrate this through identifying their strengths and weaknesses, honest self evaluation, and identifying strategies to cope with learning required or seeking support. Seeking information through asking questions and finding information independently as well as being open to feedback and coming prepared to tutorials were considered key behaviours. In addition lifelong learners were seen as participating in their own and other's learning through sharing information, reflecting on their own learning, seeking professional development opportunities, being willing to try new learning, and seeking opportunities to learn from others. It was also considered that a lifelong

learner will demonstrate change over time through their questions reflecting a developing understanding of the task, changes in performance based on feedback or personal reflection, being able to generalise learning from one context to another, and following through on learning issues/tasks as identified.

4.3. Summary

As described in the overview at the start of this chapter, the design process was reiterative and drew upon multiple sources of information and data including consultations with an expert group, students, university and field CEs, published literature, and multiple resources such as the CBOS and existing assessment tools. This chapter has summarised the nature of consultation phases of the research and the subsequent chapter will describe the synthesis of the knowledge that was developed through this process with knowledge from other sources. The considerations made in the course of determining the content, format, and process of the competency based assessment will be made explicit.

CHAPTER FIVE

5. SYNTHESIS OF ASSESSMENT DESIGN CONSIDERATIONS

As highlighted at the conclusion of the literature review, the design of the assessment format was guided by the assumptions that the assessment will occur in the workplace and has important influences upon learning. The assessment will address competent performance of speech pathology tasks and this competence requires the flexible exercise of professional judgement across a range of occupational skills as described in the CBOS and generic competencies. The validity of this assessment rests on appropriate content, format, and processes to support a valid and justifiable judgement.

Thus, from the outset, the design of the assessment was guided by the assumption that it would be used in the workplace by CEs in their dual roles both facilitating and assessing learning and that this would take the form of judging (or rating) the students' performance on both CBOS and generic competencies. Therefore, the assessment design involved determining the following:

1. Physical format of the rating scale.
2. How competence develops and demonstrated.
3. Identifying the generic competencies to be rated against in addition to the CBOS competencies.
4. Attending to the impact upon learning.
5. How to support CEs' judgement.

5.1. Format of the Rating Scale

Through reviewing the literature, consultation and investigation of other clinical assessment formats, it became apparent that there were three major decisions to be made regarding the physical format or design of the rating scale. Each decision required careful evaluation of the theory underpinning the format considerations and integration with evidence from focus groups and current practice. First, should competency be represented as a presence/absence or degree of expression of characteristics of competency i.e. progression towards competency? Second, should students be rated as to the degree they possess a characteristic proposed to be representative of different aspects of competency or should

behaviours be specified for each competency that need to be identified as observed? Third, what type of physical format should the rating scale take e.g. a VAS line, with or without markers, Likert style formats, lists of behaviours to be observed with checkboxes, or some other design?

5.1.1. Presence/Absence or Degree of Competency

A common argument within the competency based approach is that one either meets the criteria for competence or not (Harris et al., 1995). This has resulted in assessments consisting of lists of specific benchmarks indicating an acceptable level of performance and the students are judged against whether that behaviour is present or not, thus whether they are competent…or not (Curtis & Denton, 2002). Curtis & Denton (2002) argue that, when looking at generic skills such as problem solving that are applied across multiple contexts, simple benchmarks of performance are likely to be insufficient. A similar argument can be mounted when considering the complexity of professional judgement in action and the wide range of contexts in which speech pathologists exercise this judgement.

The practise of CBE in Australia has very much reflected the conceptualising of competency as 'on' or 'off'. However, evidence for the notion that there may be discriminable levels of competence rather than only competent/not competent benchmarks can even be found within CBE movement in Australia in their description of three levels of performance to be applied to all 7 of the Key (or generic) Competencies (Mayer Committee 1992, cited pp. 25 in Curtis and Denton 2002). Harris et al. (1995), in their book on CBE practice in Australia, also contend that the dichotomous approach to any competency assessment is an over simplification. They put the case that competence is not unidimensional and that a large variety of attributes contribute to final judgement of whether students are competent.

Using single benchmarks also does not address or acknowledge the importance of formative assessment which involves providing feedback and guidance to students on what to learn, how to do it, and how well they are progressing (Boud, 2000). Curtis & Denton (2002) echo this in their statement that the primary characteristics of an assessment, including those within a CBE framework, is that it facilitate the provision of feedback on the students' progress in acquiring skills and a framework for their improvement.

The notion of a continuum of competence is clearly expressed by Benner (1984) in her work on the development of expertise in nursing practice. She developed a framework of 5 categories of performance based on work by Dreyfus & Dreyfus (1996): novice; advanced

beginner; competent; proficient; expert. Benner nominated 'advanced beginner' as the level of competence required for entry into the nursing profession and suggested that the category of novice must be passed through prior to being considered competent enough to begin practise. Finally, if one considers judgement to be an integral part of professional action, the ability to make judgements has an acknowledged developmental framework of its own (Down & Hager, 1999).

When examining current practice it is apparent that all but one of the current Australian assessments of student performance in the speech pathology workplace embodies a notion of a developmental sequence of competence. The exception to this is the Clinical Assessment Form (Latrobe University) which requires CEs to nominate whether students have achieved pass level or not reached criterion at this point of assessment, however, even this assessment suggests that comments regarding the students' abilities are appropriate, implying that levels of performance prior to competency are of interest. All other assessments involve rating the student on a scale representing increasing levels of competence, with four formats nominating the end point of the scale as being the point at which students are ready to graduate, or have reached entry level competence. The exception is the Clinical Education Evaluation Form (The University of Queensland) which has the option to indicate that students are performing above entry level.

Thus it is unsurprising that both CEs' and students' discussions within the focus groups implied a developmental approach to the acquisition of competency:

> *Continuums are good – box or line, illustrates where to aim for and visually show students this.* (CE, Canberra focus group).
>
> *I am happy to comment and indicate that instead of like a tick the box reached competency, have not reached competency, more of an emerging scale of where they are at.* (CE, Latrobe University focus group).
>
> *Well, for me, having that rating scale broadened and more defined and that way you have a better understanding of exactly where you are placing within it and whether you have actually made progress or whether it has just been a tiny little shift.* (Student, Flinders University paired interview).

Other performance assessments located in the literature also have one to three levels describing above entry level performance either as a point on a likert style rating scale or as an option such as a tick box to indicate above entry level performance (Cohen, Rothman, Poldre, & Ross, 1991; Cross et al., 2001; Hrachovy et al., 2000; Johnson & Shewan, 1988;

Loomis, 1985b; O'Donohue & Wergin, 1978; Turnbull, McFadyen, van Barneveld, & Norman, 2000). This does bring into contention the issue as to whether the assessment under development should include the option of indicating above entry-level performance. All the focus groups indicated that they were in favour of placement performance attracting a nongraded pass, as is common practice in Australia. The notion of acknowledging above entry-level performance was not generally canvassed or raised.

5.1.1.1. Summary

Thus, there are strong arguments to support a developmental approach to judging competency in the hope that it prevents an overly simplistic description of the competency required, that it acknowledges that there is a developmental trajectory involved in acquiring competence, and also addresses the important formative functions of assessment. Once the argument is accepted that competence is developmental and falls somewhere between incompetent (or novice) and expert, the next issue that needs to be considered is whether the assessment of students' performance in the workplace should be described in terms of performance levels beyond competent, similar to the CBE approach.

As described above, many performance assessment tools described in the literature incorporate some sort of indication of above entry-level competence. In addition, the literature suggests that providing for indication of above entry level performance is motivating for students and educators and increases the face validity of the assessment process for employers as it is clear that, while all students are not passed into the workplaces until competent, some employees possess excellence or the potential for excellence (Harris et al., 1995; Pearce, 2001; L. Smith, 2001). However, focus groups expressed concern about the subjectivity of grading performance in workplace placements and a preference for nongraded passes. Unless the assessment tool can demonstrate that it can validly discriminate levels of performance, it should not be used to determine levels of grading or performance above a 'pass'. Overall, a developmental approach to the description of competency is supported and the notion of indicating beyond entry-level performance should be considered.

The second issue that was addressed when considering design options for the rating scale format, was how the ratings for each competency should be described. Should students be rated on each competency as to the degree they possessed of one or more characteristics or traits in relation to this competency, or should behaviours be specified that illustrate a

developmental progression on performance of that competency, or is there some other option that should be considered?

5.1.2. Rating Traits or Specified Behaviours

There are numerous types of rating scale formats represented in the literature which use a variety of approaches, sometimes in combination, to describe and judge competency. Those most commonly reported in the literature were: Visual Analogue Scales (VAS); Behaviourally Anchored Scales (BAS); Behavioural Observation Scales (BOS); Trait or Global scales; and Graphic Rating Scales. Fig. 5 illustrates how each of these might look in relation to rating performance of students on a CBOS competency.

As can be seen on this figure, VAS use simple descriptors that describe the variable in a unipolar manner, they are usually horizontal but can be vertical, and are used to measure subjective phenomena (Johnson, 1997; Wewers & Lower, 1990). The rater places a mark on the line that matches their judgement as to where the ratee's performance places them on the continuum. Graphic rating scales are very similar to VAS but have descriptors placed at intervals along a line to guide the assessor's placement of their mark (Landy & Farr, 1980; Wewers & Lower, 1990). Trait or Global Scales require the assessor to rate the assessee on a general characteristic or trait and the degree to which they possess it in relation to the competency or job requirements (Gomez-Mejia, 1988). The actual physical format of the rating scale itself can vary e.g. take the form of a graphic rating scale or likert scale as per Fig. 5.

BOS use behaviours identified as competencies required for a job, such as the CBOS, and rates them on a Likert style of scale as to how frequently the behaviour is observed e.g. almost never /almost always (Fay and Latham, 1982). BAS have anchors at various points that are described in behavioural terms and are developed through the use of a critical incident technique that involves a consensus process whereby ratees and/or raters determine the dimensions to be rated. The end result is usually a scale with behaviours for differing levels of performance for a particular dimension ranked vertically for the rater to select and these are assigned a scale value (Barnhardt, no date; Fay & Latham, 1982; Landy & Farr, 1980).

Figure 5. Illustration of rating scale formats using CBOS Competency Unit 3 as an exemplar

---

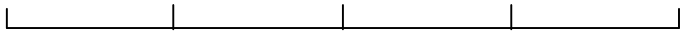
Example competency for rating:

CBOS Competency Unit 3: Planning of Speech Pathology Intervention

Visual Analogue Scale

Poor Excellent

Graphic rating scale



1 2 3 4 5

Poor Average Excellent

Global Trait Scale

Interdependence

1	2	3	4	5

Creativity

1	2	3	4	5

Organisation

1	2	3	4	5

Behavioural Observation Rating Scale

Element 3.1 Uses integrated and interpreted information (outlined in Unit 2) relevant to the communication and/or swallowing issues, and/or the service provider's goals to plan speech pathology intervention.

Almost Never 1 2 3 4 5 Almost Always

Element 3.2 Seeks additional information required to plan speech pathology intervention.

Almost Never 1 2 3 4 5 Almost Always

Element 3.3 Discusses long-term outcomes and decides, in consultation with client, whether or not speech pathology strategies are appropriate and/or required.

Almost Never 1 2 3 4 5 Almost Always

Element 3.4 Selects speech pathology program or intervention in conjunction with the client and significant others.

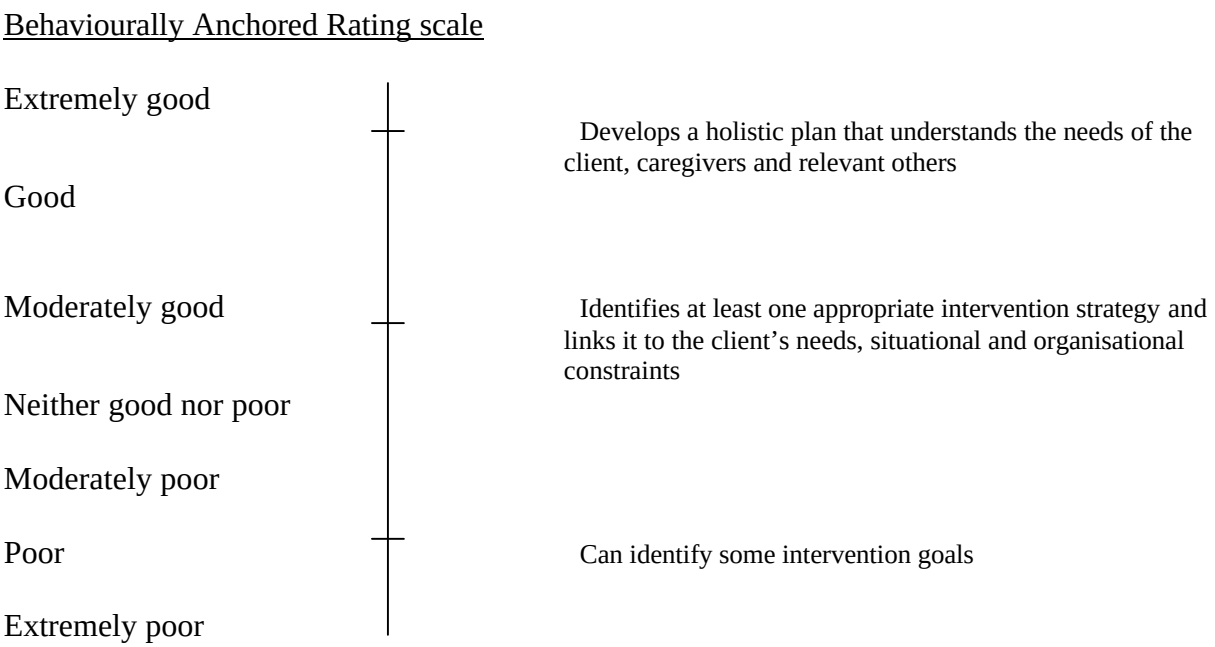
Almost Never 1 2 3 4 5 Almost Always

Element 3.5 Establishes goals for intervention.

Almost Never 1 2 3 4 5 Almost Always

Element 3.6 Defines roles and responsibilities for the management of the client's swallowing and/or communication condition and issues.

Almost Never 1 2 3 4 5 Almost Always



As can be seen from this description, these scales are not independent in terms of format and content and are sometimes used interchangeably or quite idiosyncratically and all could be equally applied to the rating of speech pathology competencies. For example the competency could be rated against the presence/absence or degree of expression of various characteristics, either combined into one concept of important aspects of competency, or separated into several different components. This rating could be reflected as a present/absent dichotomy, or as various degrees marked on a VAS line or a Likert style line with numbers placed along it. Alternatively the behaviours that represent presence/absence or degree of expression of the competency could be detailed either as a list with checkboxes, or placed along a VAS line e.g. to reflect increasing mastery of the competence.

These rating scales can be differentiated in terms of the way in which they organise the information on which students are to be rated or the way in which the judgement is to be made. BAS and BES scales are designed so that students are rated/judged according to a list of behaviours specified within the rating scale. Alternatively, BOS scales are structured so that the behaviours of interest are listed and the assessee is rated as to how frequently these behaviours are observed. All of these approaches generally involve specifying the behaviours to be observed in some detail. The other option is to rate the assessee according to the degree to which they possess a trait in relation to particular competencies e.g. Trait/Global, VAS or Graphic rating scales.

5.1.2.1. Current Practice in Speech Pathology and Other Health Sciences

The most common practice in Australian speech pathology competency assessment has been to rate the students as to the degree to which they are judged to possess one or more traits in relation to each competency e.g. independence, creativity. Some of these traits are described in terms of how they would be observed behaviourally but in the majority of cases CEs are largely required to interpret how the trait would be expressed in the students' behaviour in relation to particular CBOS competencies. Thus a combination of ratings of traits and specifying behaviours deemed to express these traits is usually employed. The Clinical Education Evaluation Form (The University of Queensland) is somewhat different to formats used by other universities as it uses a BES approach. In this case, the rating format comprises a continuous line divided into segments, each segment identified by a number, and a list of behaviours representative of that rating number in relation to that specific competency. The Clinical Protocols (Flinders University) also differs, in that it uses a form that rates the students according to how much they possess a trait described as 'independence' but the areas they are rated against could be described more as behaviours than competencies.

This variety of approaches is also reflected in the literature. Hrachovy et al. (2000) ask CEs to judge physiotherapy students against a list of skills, each skill being described by key indicators, and to select between four criteria to describe this judgement: exceeds entry level; meets entry level; needs experience to meet entry level; needs improvement to meet entry level. Other approaches take a more global approach and rate the students against each behaviour or competency according to how well they were judged as performing e.g. 'low' to 'high' performance (Cohen et al., 1991). Roach et al. (2002) provide a list of skills and ask the CEs to express their judgement of how students' behaviour expresses a combination of traits and other considerations as a single mark on a VAS. The traits to be considered are consistency of performance, efficiency of performance, and supervision/guidance required, in combination with judgements as to the overall quality of care provided, all referenced against the complexity of tasks/environment the students have experienced.

The most common approach reported in the health science literature is a combination of checklists of specified behaviours with assessment under controlled conditions e.g. Luttrell et al. (1999) assessment of nursing students' competency an assessment or OSCE style assessments of performance as described in the literature review. More recently, global rating(s) have been added to OSCE checklists where the examiners are asked to make an

overall judgement of the students' performance on the observed tasks (Cohen et al., 1991; Friedman & Mennin, 1991; Norman, van der Vleuten, & de Graaff, 1991).

As identified above, many performance assessments do require the assessor to judge a combination of traits and/or general characteristics of performance sometimes in combination with specific aspects of behaviour. The difficulties this can pose for both assessors and assessees is reflected by the consistent plea by all focus group participants for greater clarity in descriptions and more specific examples of behaviours on which to base judgements.

> *I would want it to be self-explanatory. Not too ambiguous. That each of the points that you are actually rating make sense to the educator and the student you don't end up explaining why you, what you are scoring.* (CE, Adelaide focus group)
>
> *I am far more concerned about the descriptors than the numbers. If the descriptors were more accurate it would be easier to plot the level of competence on any scale. It's just that description of what you measuring that I find very difficult.* (CE, The University of Sydney focus group)
>
> *Clearer wording need to discern between each level, what the actual meaning of the statements are.* (CE, Canberra focus group).
>
> *I think you need more…probably a series of meetings and criteria to be able to move onto the next level. But I don't know how do you know what that criteria is?* (Student, The University of Newcastle interview)
>
> *I know that there's a sheet that tries to define each level, but I haven't found that particularly helpful. I still find it hard to work out exactly where I would be.* (Student, Flinders University paired interview)

This point of view is supported by Loomis (1985a) who reviewed the medical education literature and found that the consensus was that, to improve rater reliabilities, competencies and associated performance standards need to be well defined in terms of observable behaviours or standards that describe the levels of mastery of the competency.

5.1.2.2. Evidence From Assessment of Job Performance in Other Fields

The effectiveness of different rating formats has been evaluated in relation to assessing job performance for recruitment or promotion. BAS have received the most attention with advantages suggested to be primarily due to the fact that the rating categories are descriptive, development involves the raters and ratees and thus improves validity, they are specific to the

job and concrete rather than global and abstract, and introduce less construct irrelevant variance (Fay & Latham, 1982; Gomez-Mejia, 1988; Kingstrom & Bass, 1981).

However these proposed advantages are not supported by empirical evidence. For example Kingstrom & Bass (1981) conducted a detailed review of 21 studies that compared BAS with other types of rating scales and found that the BAS were not superior to alternative scale formats. Overall, the evidence suggested that the format of the scale did not affect its ability to sufficiently differentiate between different levels of performance and that BAS were not necessarily perceived more favourably than scales developed through different processes. Gomez-Mejia (1988) compared BAS scales for assessing the work performance of 219 technicians with ratings on two different global dimensions (creativity/attitude and potential for promotion) and found that the BAS was not necessarily superior. In fact, the global scales showed slightly greater criterion related validity and generally lower interscale correlations than did the BAS, with a lower halo effect. Gomez-Mejia (1988) suggested that this may be due to the BAS task requiring too may specific judgements in the perception of complex and numerous job behaviours and thus overburdening the rater and resulting in their judgements becoming less discriminating. This may also be due to the assessor placing a check against a behavioural anchor that does not really represent the ratee's performance on the job in the absence of a better descriptor (Fay & Latham, 1982; Wolfe & Gitomer, 2001).

Wolfe & Gitomer (2001) avoided this effect when they altered the scoring rubrics for an essay assessment so that the examiner's attention was directed to observed qualities of performance rather than specific behaviour. Evidence was found that indicated that this change improved the validity of their assessment process. However, it should be noted that aspects of the rater training process were also modified.

Landy & Farr (1980) reviewed the literature published between 1950 and 1980 specifically on performance rating and did find that scale anchors are important and that there is some evidence to suggest that behavioural anchors are better than numerical or adjectival ones, particularly if the dimensions being rated are poorly defined. This advantage is suggested to relate to behavioural descriptors ensuring that the rater has a clear understanding of the rating task and that rigorously developed anchors, that are more than simple descriptive labels such as poor/average/excellent, support their judgement more effectively. Overall, Landy & Farr propose that superior scales may simply be the result of psychometric rigour in their development and of some level of participation by the individuals representative of those who will eventually use the scales to rate, rather than some characteristic unique to

behavioural anchors. Thus rigorous item selection and anchoring procedures are important regardless of what type of scale is used (Landy & Farr, 1980).

Curtis & Denton (2002), developed an assessment of problem solving that described a hierarchy of observed qualities of performance specific to each sub competency for the domain of problem solving and identified how well the skills of problem solving are expressed in student performance. This assessment used evidence from the literature to identify the components and the developmental sequence of problem solving behaviours within each sub competency and was found to have strong validity qualities. However, as has been mentioned before, there is currently very little evidence based information regarding the process of speech pathology work that can be used as a basis for this kind of rating approach.

5.1.2.3. Summary

The literature on performance assessment constantly emphasises that "principles for the design of performance assessments are very much in their infancy" (pp. 91, Wolfe & Gitomer, 2001). However, current practice, research and the focus group consultations suggest that the following practices should be consider in order to maximise validity:

1. Use global ratings of qualities of performance rather than checklists of specific types of task behaviours.
2. Provide clear descriptions of how these qualities may be evidenced in behaviour.
3. Use anchors that are related to behaviours of interest and rigorously developed.
4. Ensure that the rating tasks do not require too many complex and numerous specific judgements.
5. Involve people who are representative of both raters and ratees in the development of the rating task.
6. Ensure that the entire scale development process is psychometrically rigorous.

The next issue to be addressed is whether there are any scale formats that have particular advantages for the assessment tool under design.

5.1.3. Influence of Scale Formats

There is evidence within the literature that different assessment methods e.g. multiple choice, portfolios, rating workplace performance, result in poorly correlated measures of student performance and thus appear to sample different components of student performance

(Newble & Swanson, 1988; O'Donohue & Wergin, 1978; Shavelson et al., 1993). However, once it is determined that a performance assessment is required, reviews of literature relevant to performance ratings suggest that format of the rating scale may not be as critical as other factors. As mentioned previously, Kingstrom & Bass's (1981) review of research comparing BAS and other scale formats including VAS, Graphic Rating Scales, Likert styles, or mixed formats, found that there was little or no difference between them in terms of psychometric characteristics. Landy and Farr's (1980) review of performance highlighted that only 4 to 8% of variance can be explained on the basis of format and that this may even be an overestimate.

Thus the question becomes whether there are arguments in favour of a particular rating scale format that may contribute to the validity of the assessment format. Current practice by 7 of the 8 Australian speech pathology programs is to use scales that represent a continuum of development. Six of these are adaptations of a Likert style rating line and require selection of a particular number (usually from 1 to 5) to represent the judgement of competency.

Discussions with the expert reference group led to the researcher evaluating the use of a VAS style rating format. This type of rating scale was used by the American Physical Therapist Association (APTA) (Roach et al., 2002) for use by American University programs. The reasons cited in this research for favouring a VAS included that a VAS was more appropriate for evaluating complex human performance that cannot/shouldn't be divided into discrete units of behaviour and reflects degree of change better than categorical scales. VAS scales were also suggested to avoid issues such as end aversion bias which decreases the actual number of points considered in the assessment e.g. reduces 5 point scales to 3 points, people adding plus or minus designations or decimals to categorical scales, and numbers which can have unintended meanings attached in the assessment context.

The expert reference group also suggested that a VAS would facilitate individual universities developing their own grading systems if these are required. This could be done by identifying the rating ranges based on measurements from 0 to 100 on the VAS scale, that represent various judged levels of performance. In addition, even if grading was not required, the measurements may enable designation of appropriate cut off scores to represent failing performances. The option of a VAS format was evaluated through reference to published research, the focus group data, and seeking further consultations with CEs and students.

5.1.3.1. Clinical Educator and Student Opinion re VAS

As mentioned previously, CEs and students generally expressed a preference for a scale that represents a continuum of development and the sensitivity of the assessment tool to changes in the students' level of competence across a placement and over the course of a program was a prime concern for all. This suggested a VAS might be appropriate. Most, with the exception of the Northern Territory CE focus group, did not actually specify that it should include numbers and boxes or a line seemed equally preferred. In fact, many indicated a dislike for numbers currently used on scales. Issues included that there never seemed enough numbers to represent small increments of progress effectively and CEs in particular were concerned that they may encourage arbitrary ratings based on what number is assumed to be appropriate for students at a particular point in their practicum experiences, rather than their actual performance. Specifics regarding preferred rating formats were only mentioned by two focus groups, The University of Newcastle and The University of Sydney CEs, both of whom discussed a VAS like idea.

Scale design options were presented to a group of twelve CEs associated with The University of Sydney and a second group of 10 second and third year students enrolled at The University of Sydney (Appendix 12). The CEs primarily preferred a VAS scale, with the words 'novice', 'intermediate' and 'entry-level' placed along the line at the start, centre, and end, with some CEs not requiring 'intermediate' to be designated. A few people still preferred a numbered scale.

The student group was concerned that the VAS was too arbitrary as to where they would be marked on the line and whether a very small and possibly chance variation in the point at which the line was marked would have a disproportionate effect upon their assessment result. Overall they preferred the scale format similar to a VAS but with vertical marks placed along the line to indicate five stages of development, with novice, intermediate, and entry-level points being labelled. Students also highlighted that clear statements should be made about what each point on the line represented. In general students preferred having five points identified on a continuum to having numbers or a VAS scale. Overall, this consultation suggested that a modified version of a VAS format might be acceptable to both groups.

5.1.3.2. Evidence in the Literature re VAS

The VAS style rating used by the tool developed by APTA has been critiqued as being difficult to interpret by assessors because it has no anchoring descriptors on the line itself, and it was difficult to determine how much change in the VAS rating represents a noticeable improvement (Hrachovy et al., 2000). There appears to be no research comparing VAS formats to other rating scale formats and it is assumed on the basis of reviews by Landy & Farr (1980) and Kingstrom (1981) that the scale format is not critical. However, Wewers & Lower (1990) critically reviewed the advantages and disadvantages of VAS in relation to measuring clinical phenomena such as pain or mood, and found that VAS were a convenient, easy, quickly administered measurement for subjective phenomena. However, they identified that clear and careful definition of the phenomenon to be measured is required and that VAS are often used to measure a multidimensional construct with a unidimensional format, so it can be impossible to identify which dimension is being evaluated by the subject. Their review highlighted that the level of measurement needs to be considered (ordinal or interval) and the distribution of VAS scores evaluated.

The majority of these issues have been addressed earlier as they serve multiple needs. However, it is the fifth issue above, regarding level of measurement and how scores are distributed along the VAS that alludes to some critical issues. First, is the data from a VAS in fact continuous (Linacre, 1998b)? Can raters truly differentiate between an infinite number of choices on a continuous scale? Linacre (1998b) and Munshi (1990) both point out that assumptions around Likert and VAS type scales have rarely been empirically tested, resulting in practices that are unsupported by evidence. Linacre (1998b) cites two reports regarding the use of VAS for rating (Munshi, 1990; Thomee, Grimby, Wright, & Linacre, 1995) and highlights that, through the use of Rasch analysis (a statistical approach to evaluating data that will be described later in this thesis), it could be identified that both scales used in this research were in fact used by respondents as if they were categorical.

So, for example, a measurement of 98 could not necessarily be demonstrated to represent a different judgement from the measurement of 99. Thomee et al. (1995) found that the VAS scale used actually contained only 10 replicable category groupings as opposed to the 101 categories obtained by measuring the 10 cm line into 1 mm intervals. Even with only 10 categories, it appeared from the analysis that the extreme categories (1, 9, 10) were being influenced by idiosyncratic responses to pain (Thomee et al., 1995).

Munshi's (1990) research actually set out to develop a method to empirically determine the number of choices that should be used on a Likert scale and to investigate whether the Likert assumption of equal intervals and symmetry along the rating scale held true on practice. The strategy used by Munshi (1990) was to ask subjects to indicate strength of agreement regarding statements related to satisfaction with airline travel on a VAS scale. He then subsequently analysed the scale through cluster analysis to determine how the scale was used by respondents and found that the scale was very symmetrical but that the distances between scale points were not equal. Munshi found that the distance between complete agreement and strong agreement was one third of the distance between simple agreement and strong agreement. In addition, using 4 categories could cover 75% of the variance and using 7 categories explained 98% of the variance, as opposed to more detailed measurement intervals.

Munshi's (1990) approach is particularly interesting as he states that the quality of a scale is determined by its ability to faithfully reflect the attitude or opinion to be measured. He proposes that a response scale can be constructed that closely matches the internal response of the rater. This reduces measurement error that arises when a rater is forced to express an opinion via an inadequate scale. This approach is congruent with Michell's (1997) argument that psychometric assessment should ensure both that the characteristic being measured is truly quantitative as well as constructing a procedure to estimate, as reliably as possible, the quantities of the variable of interest held by the persons being assessed. Michell suggests that measurement in social sciences is "in the grip of some kind of thought disorder" (pp. 355) as it persists in an inadequate definition of measurement, i.e. that measurement consists of assigning numbers according to a rule, and then treating these numbers as having the same properties as other kinds of physical measures, such as height, without actually testing this assumption first.

In other words, the data yielded by measuring the VAS can not necessarily be accurately represented as equal and increasing amounts or quantities of a particular characteristic and the numbers treated as such e.g. through statistical analyses based on the assumptions that the data is interval. In fact, data from a VAS may be more appropriately represented by categories and it cannot be assumed that precision of a scale increases (thereby decreasing the degree of measurement error) as the number of choices is increased. Indeed, a greater degree of measurement error may be introduced by the 'noise' created by having too many categories. Linacre (1998) argues that reducing the categories on the VAS has not lost any replicable information and this ensures that statistics are calculated on the basis that the data is what is says it is. In fact, Linacre and Landy & Farr (1980) both cite research by Miller, published in

1956, which suggests that seven levels, plus or minus two, are the finest degrees of perceptual discrimination humans can make in any situation. Thus, having a greater number of measures from a scale may be redundant and misleading.

However, it does appear that statistical analysis of VAS measurements may assist in developing an understanding of how the phenomenon of competence is quantified by raters and perceived to develop. Thus starting with a VAS scale and evaluating its measurement properties may result in a more valid assessment. Landy & Farr (1980) essentially suggest a similar approach, i.e. investigating how a scale is used to evaluate the appropriateness of the scale format for the task, when they recommend that, if a continuous rather than discrete response continuum is being considered, pilot studies should be run to determine how many response categories are perceived by the potential raters. Linacre (1998b) suggests that Rasch analysis ideally lends itself to the task of analysing how many categories operate in reality on VAS scale and it would appear that cluster analysis is also worthy of consideration for this task (Munshi, 1990). Determining how a rating scale is interpreted by the raters, rather than assuming it's meaning, prior to using that data to make decisions regarding students' competence has appeal and would greatly support its validity, particularly its consequential validity. Furthermore, this process would very much contribute to the structural validity of the assessment tool, which is described by Messick (1989; 1994; 1996) as the fidelity of the scoring structure, as it would be congruent with the structure of the construct domain being assessed.

5.1.3.3. Summary

There does not seem to be sufficient evidence in the literature to entirely support all the arguments made by the developers of the APTA assessment tool regarding the superiority of the VAS for performance rating (Roach et al., 2002). It is also not clear whether the VAS performs better with respect to end aversion bias as suggested by Roach et al. (2002) as they do not report on this issue in their research. In addition, Thomee et al. (1995) found that some end aversion may have been occurring with their VAS scale. However, given that that it is proposed to have entry-level as the end point of the rating scale (with the option of indicating above-entry level performance), it seemed unlikely that end aversion bias would be a strong influence on rating behaviour at least for the top end of the scale. The literature and focus group data does support that it is preferable to not use numbers on a performance rating scale, suggesting that a VAS may be a better choice.

It seems likely that the data derived from measuring a VAS is not as continuous as suggested by Roach et al. and may in fact be more categorical than interval. On the other hand, using a VAS may provide an opportunity to derive information from the ratings that can be used to safeguard the structural validity of the assessment and it seemed likely that statistical approaches were available that would provide information on how the VAS is used in actual practice and what kind of data it yields. This may also provide information to determine if students' concerns regarding the arbitrariness of markings on a VAS line are valid. Furthermore, the concept of an adapted VAS scale was preferred by a group of CEs and students who had an opportunity to review the scale design options. Finally, focus group data was either neutral or supported the notion that a VAS may be appropriate for assessing student performance in workplace settings and all groups were definitely seeking a scale that would better reflect even small amounts of progress.

Thus it was decided to use a VAS or graphic rating scale (line with some divisions on it) in the first instance and to analyse the ratings to determine whether a VAS or categorical scale better represented the data. A VAS could in fact continue to be used on the assessment tool as this analysis process would simply determine how measurements from the VAS should be interpreted with regard to decisions that rely on this data e.g. pass/fail or grading performances.

5.2. Identifying Levels of Competence

As has already been argued, a developmental approach to the assessment of competency is supported on the basis of focus group, expert opinion, current practice, and the literature. Focus group opinion, as well as consultations at the SPAA 2002 conference, suggested that it would be preferred for each competency to be rated on one dimension and that this dimension should be consistent throughout the assessment.

This led to consideration of two closely related issues with respect to the final design of the rating scale. First, what dimension or aspect of a person's performance on a particular skill should be rated to determine their competency and should this dimension be the same for the CBOS items as for the items representing underpinning skills? Second, what does the progression from a novice to competent or entry-level practitioner in speech-language pathology involve? This is an important factor in determining how to represent progression along the dimension being measured.

Again, multiple sources of evidence were examined including current practice, as reported in the literature and evidenced in Australian assessment formats, and theoretical understandings regarding the transition along the continuum towards competence and possibly beyond.

5.2.1. Ratings of Competence Reported in the Literature

The literature on how students' competence is assessed is somewhat frustrating in that it does not often identify the skill that is being rated or what is represented on the low to high points on the rating scale. For example, Dauphinee (1995) who reviews assessment of clinical performance in medicine and even critiques the use of ratings in these assessments, does not actually identify what dimensions students are rated on. Thus the literature yields little specific information to guide decisions regarding how levels of development can be defined.

Many of the rating scales described indicate that the assessor rates each student's competency as a matter of degree. For example, Stackhouse & Furnham (1983) refer to ratings of speech language pathology students' competency on a 7 point rating scale across the range of inadequate, poor, adequate, satisfactory, good, very good, and excellent. Cross et al. (2001) explore the use of ratings of video vignettes of physiotherapy students as an assessment of their competencies in clinical situations and describe a 10 point scale where 0 is unacceptable performance and 10 exceptional performance, with 4 being the pass/fail cut off point. Both of these studies identify some sort of progression over a dimension of performance or competence that is not made explicit.

Norman et al. (1991) review the pros and cons of various methods of assessing the practical skills of medical students with regard to reliability and do identify the aspects of the performances being rated on an OSCE format as including specific ratings on the students' technique with regard to the skills involved, their fluency in using the technique, and the quality of their approach with the patient. However, how the examiner determines performance between the lower to upper ends of these scales is not described. Similarly, Hrachovy et al. (2000) describe the tool they use in assessing physical therapy students'

clinical competence, The Blue MACS (5th Edition), as requiring a rating of the students' progression on their performance of particular skills over the following trajectory:

1. No ratings i.e. haven't been working on it.
2. Needs improvement to meet entry level.
3. Needs experience to meet entry level.
4. Meets entry level.
5. Exceeds entry level.

Key indicators of entry-level mastery are provided, but it is not clear if the other anchor points of the scale are described.

The APTA scale (Roach et al., 2002) does provide more information on how assessors are directed to rate students' performance on 24 skills using a VAS with reference to the end point criterion of 'entry-level performance'. The decision as to where to mark the VAS is a judgement that results from considering five dimensions of performance: quality; supervision/guidance required; consistency; complexity of tasks/environment; and efficiency. Each of these dimensions is briefly described in terms of a low and high level of performance, with some information regarding how performance may change on these dimensions over the placement, as well as novice and clinical performance.

5.2.2. Current Australian Conceptualisation of Dimensions to Rate Regarding Student Speech Pathology Performance in the Workplace

5.2.2.1. Current Australian Assessments

As has already been described, the performance dimension(s) of the skill to be assessed in the workplace and how progression over this dimension(s) is identified varies between Australian universities currently offering speech pathology programs. Most commonly, students are rated on a scale of 5 or 7 points as to the degree to which they demonstrate one or more traits in relation to a competency. All formats have, of course, attempted to define in general terms the developmental progression to be considered and how students' performance will generally change as they progress in competency along the continuum described by the dimension. However, focus groups strongly indicated that this information needed to be more explicit or perhaps presented in a different format to assist judgment of performance on these continuums.

Table 8 illustrates the variety of approaches currently used in Australia to address the complex task of determining the competency levels of speech pathology students' performance in the work placements. There are several aspects to consider when reviewing the way in which the development of competence in speech pathology students is currently conceptualised in Australia.

All the assessment tools currently in use acknowledge either directly or indirectly the skills outlined in the CBOS but also, in an apparent attempt to acknowledge generic competencies, assess a combination of other skills in relation to the CBOS skills. They approach this in a number of different ways. CBOS skills may be rated on several different performance dimensions or global ratings included that assess performance across the CBOS skills in total. Both these performance dimensions and separate sets of global ratings could in turn be conceptualised as generic competencies that have several components to be addressed. Other assessments rate competence on generic skills in which the CBOS skills are implicit or the assessor is directed to the sections of the CBOS covered by the particular generic skill described. Some Australian assessment tools rate the CBOS competencies on performance dimensions that could perhaps be better described as competencies themselves and are similar to the concept of generic competencies described in the literature review. In general, few of the tools base their taxonomies on a theoretical model of developing professional competence, with the exception of the 'interdependence' concept owing its origins in work by Brasseur (1989) and 'independence' related to notions of decreasing scaffolding for learning by Bruner (1983).

Table 8. Aspects Rated in Assessments in Current Australian Assessments of Speech Pathology Students in the Workplace

Dimension of competency to rate	Description of progression towards competency
Type of Competency: General Occupational Competencies with CBOS implicit	
Increasing complexity	Specific and comprehensive listing of the types of behaviours that would be observed for each skill at each level of the progression towards competence in that skill.
Independence	Change in the amount of supervision students need to be competent i.e. progressing from observing only through to working without direct supervision, with change described in terms of the amount of scaffolding by the clinical educator required to support the students' performance
Type of Competency: Generic Competency	
Professionalism	Change over a number of dimensions e.g. responsiveness to the needs of the client and the service, responsibility for planning and service delivery, ethics and legal requirements, commitment to quality service
Interdependent learner	Change over a number of dimensions e.g. awareness of learning needs, responsibility for own learning, type of interaction with the clinical educator that is used to seek information
Adaptability and creative thinking	Change over a number dimensions e.g. identifying need for change in client or their own behaviour and timing of the response to this (delayed to immediate)
Self-evaluation	Change over a number dimensions e.g. evaluating their own or clients' behaviour with a shift in focus from the students' to the clients' needs
Type of Competency: CBOS	
Interdependence	Progression through the following general behaviours: moving from constant assistance and direction through to occasional supervision and sought at students' own initiative
Independence	Degree and type of supervision required for students to be competent
Collaboration	Progression through the following general behaviours: support and direction from supervisor; through to works with peers to enhance learning opportunities and benefit client management
Complexity	Progression through the following general behaviours: manages parts of simple, straightforward individual cases; through to manages total, complex caseload demands
Efficiency	Progression through the following general behaviours: time taken and quality relies on clinical educator input; through to work is of optimum quality given time available and caseload demands
Critical/Creative Thinking	Progression through the following general behaviours: follows prescribed procedures and can reflect upon and evaluate own clinical performance; through to develops and implements innovative clinical procedures in the light of critical evaluation of the research literature in the field of practice

5.2.2.2. Focus Group Data and Australian Literature

The focus group data suggested that assessment tools that rate the CBOS skills on several different dimensions are difficult to use.

> *Huge grids or matrixes are difficult.* (Field CE, Canberra focus group)
> *And also some of the areas that you are assessing them on there are several aspects in the one mark you are giving, and in some of them they might be doing quite well but in other aspects of the same thing they might not be doing so well in. And the, it's which part do I, what sort of mark do I give them?* (Field CE, Northern Territory focus group)
> *Just one dimension would be easier.* (CE, The University of Newcastle focus group [field note])

It seems likely that making judgements on each CBOS competency, by referring to 3 or 4 different dimensions of skills (generally referred to in the focus groups as grids or matrixes), that in fact appear to be generic competencies in themselves (e.g. professionalism, critical/creative thinking), is particularly difficult. This indicates that it would be helpful to attempt to describe a continuum on which to rate students as a single developmental progression on each specific competency being assessed (occupational or generic) rather than confounding assessment of one competency with another. This suggestion was also supported by the comment that the requirement to make several different discriminations in relation to one competency and then combine them into one rating may have contributed to moderate rather than high inter rater reliability on the APTA tool (P. Hagler, personal communication, September 2002).

In addition, CEs and students indicated that, for them, markers of developing competency included increasing ability to manage complexity, the increasing ability to focus on the client and not themselves or the task at hand, and greater confidence. For example, field notes from the SPAA conference discussions identified the following as the single dimension on which to rate students:

1. Adaptability in attitude, skills, and knowledge.
2. Isolated practise (e.g. simple tasks) to integrated practise (complex tasks) e.g. understanding, synthesis, integration, application.
3. Self focus to client focus i.e. context focus, bigger picture focus.

Flinders University students in one focus group interview suggested that their development of competence should be judged in terms of their development of confidence in themselves and their ability to learn, a CE in the Adelaide focus group summarised the CE's perspective on this as "*It comes back to that confidence versus knowledge of their limitations.*" This notion of independence and interdependence was referred to frequently across all focus groups.

A number of these themes were evident in the categorisation of novice, advanced beginner, and entry level skills for students by McAllister and Lincoln (2004) who identified changes in affective components, automaticity, and focus. The affective components of students' performance include decreasing anxiety and improving confidence. Automaticity relates to multiple aspects of performance including using clinical reasoning and making decisions, organising time and sessions with clients, and improving ability to tolerate and manage complexity. Changes in focus relates to students' moving from focussing on themselves and the session towards attending to the whole client and situation, using information from a range of contexts and people for assessment and intervention. While these were perceived as different components of increasing generic competency they can be usefully classified as aspects of an overall ability to manage complexity effectively.

5.2.2.3. Summary

The dimensions or aspects of speech pathology students' performance currently measured in Australian workplace assessments are many and varied. There is no clear consensus on one or even a few specific dimensions that should be used to rate students' competency on assessment items. There also appears to be no theoretical underpinning to the progressions on the rating scales used to map out the development from novice to competent, with the exception of concepts of independence or interdependence. A consistent theme related to students' ability to effectively manage complexity within the workplace is also evident. A means for categorising or describing the progression of students from novice to competent or entry-level practitioners would greatly assist development of items to describe this progress. This would facilitate clear decision making as to how increasing competence on a particular skill will be reflected in the behaviour of the student.

5.2.3. Describing the Progression From Beginner to Competent

The specificity and amount of the descriptors used to illustrate progression on competencies defined in Australian assessments of speech pathology workplace performance varies. The Clinical Education Evaluation Form (The University of Queensland) has the most exhaustive lists of behaviours and provides them for each of up to 7 rating points on their scale. Other tools that rate some 'generic' competencies, e.g. flexibility, provide brief descriptions of general behaviours that illustrate levels of competency across 3 anchor points on a 5 point scale. Some assessment formats, such as the Clinical Protocols (Flinders University), outline the characteristics that illustrate 5 different levels of performance and require CEs to apply and interpret these with reference to each of the competencies within the assessment.

The validity of these descriptors would generally appear to be strong as all have been refined and developed over time through cycles of expert consideration and consultation by speech pathology programs. However, the majority do not describe a theoretical framework and as such offer few clues as to strategies for analysing and developing descriptors for the competencies represented in the revised CBOS (2001) or Generic Competencies section of the proposed tool.

Notions of independence or interdependence are the exception to this, relating to theories espoused by Bruner (1983), Anderson (1988) and Brasseur (1989). However, it could be argued that independence / interdependence is in itself a dimension that students should be rated against and indeed has been conceptualised this way in some tools. It is also not clear that developing interdependence is an appropriate or, indeed, the only relevant marker of increasing competence on a particular skill. However, it is a concept that CEs seem to be at ease with, for example:

> *I think level of independence is a really important one.* (CE, Northern Territory focus group)

However, this concept can be misinterpreted in application, for example:

> *The independence scale is a nice idea but it doesn't work. Everyone has got different ideas on what it means.* (Student, Flinders University paired interview [field notes])
>
> *…because you are being rated on independence you almost feel like if you are going to ask the supervisor things, you are being too dependent and therefore you are going to receive a lower score.* (Student, Flinders University paired interview)

There are a few theories in the literature that describe or explain what is encompassed in the transition from being a 'beginner' clinician through to being 'competent' or even 'expert', or (as mentioned previously) even what actually occurs during health professional practice. Some of these theories address limited areas of practice, for example cognitive theories of knowledge acquisition, and are usually applied to clinical reasoning in relation to assessment, diagnosis, and treatment planning in medicine (Boshuizen & Schmidt, 2000; Newble, van der Vleuten, & Norman, 1995). These clinical reasoning theories provide useful insight into what health professionals believe they are doing when working effectively with clients/patients and many of these skills can be described as generic. Unfortunately they do not describe a developmental pathway that could be used to identify how students are progressing in general along the continuum from beginner to competent. Other authors do not detail a developmental pathway of competence but detail their understandings of what comprises professional expertise, such as Higgs & Bithell's (2001) description of the dimensions of expertise. However, a few models were identified in the literature that appeared to be useful starting points for defining a developmental progression towards competence against which to rate student performance.

5.2.3.1. Stages in Skill Acquisition

Dreyfus & Dreyfus (1996) have developed a model, originally in the late 1970's, to describe the progressive development of skills from beginner through to expert. This model was used by Benner (1984) in her analysis of the development of competencies in nursing practice. Dreyfus & Dreyfus (1996) identify that, as people improve their skills and knowledge through experience, they move through 5 stages of qualitatively different perceptions of their task. This development is dependent on the reciprocal interaction of theory and practice, with one informing the other. Dreyfus & Dreyfus assert that expert judgement of what is required in a situation evolves from this experience and is more than the application of theory-based principles as it also involves the application of intuition derived from experience. Benner, Tanner, & Chesla (1996) in fact define 'intuition' as a judgement made without first considering a rationale, it is an immediate understanding of a clinical situation which is then assessed using more deliberative, analytical, or logical thought processes. This complex and reciprocal intertwining of theory, practice, and judgement (both explicit and intuitive) to determine decision making and action explains why experts outperform computer programs based on applying principles to situations (Dreyfus & Dreyfus, 1996). These concepts are of course very congruent with the notion of professional

competence involving the competent exercise of complex professional judgement across all tasks and contexts of the profession, as argued in the literature review.

Table 9. Stages of Development of Progression From Novice to Expert (summarised from Benner, 1984; Dreyfus & Dreyfus, 1996)

Developmental Level	Characteristics
Novice	• Uses limited, inflexible rule governed behaviour derived from principles and theory they have been taught prior to experience in the real situation • Are unable to identify the most relevant task or issue in the actual situation
Advanced Beginner	• Marginally acceptable performance • Able to notice the recurring meaningful aspects of the situation and now has guidelines or principles that dictate actions in terms of attributes and aspects of a situation • Takes in too little of a situation, because they have to concentrate on remembering the rules that they have been taught. • Has difficulty prioritising, as all aspects of a situation appear to be equally important
Competent	• Now able to see their actions in terms of long range goals or plans, and are consciously aware of these • Plans dictate what attributes or aspects of a current or future situation are most important and what can be ignored • Able to establish a perspective based on considerable conscious, abstract and analytic consideration of the problem • Not as fast or flexible as a proficient nurse but do have a sense of mastery and ability to cope • Rely on conscious planning to assist their efficiency and organization
Proficient	• Now sees a situation as a whole • Performance is guided by maxims that would not make sense to a less experienced practitioner • Able to recognise what is important based on previous experience rather than conscious thinking • Will shift to an analytic or competent level approach in novel situations
Expert	• Does not use analytic principles to connect an understanding of a situation to an appropriate action • Has an intuitive grasp of each situation and performance becomes fluid, flexible and highly proficient e.g.’ have’ a feel for what they are doing • Have highly skilled analytic abilities and applies these to new situations or when they recognise that events are not occurring as expected

There are changes in 3 general components of skilled performance which result in the development of expertise across 5 stages as described by Dreyfus & Dreyfus (1996) and Benner (Benner, 1984). First is a movement from reliance on abstract principles to use of past concrete experience as paradigms. The second dimension regards change in learner's perceptions of the situation in which the situation is seen less as a compilation of equally relevant bits and more as a complete whole in which only certain parts are relevant. Third, with developing expertise the practitioner moves from detached observer to involved performer. These changes, in combination, result in five stages in the development of expertise summarised in Table 9.

Unfortunately, in Benner's original work (1984) where she identifies nursing competencies and skills similar to the competencies described CBOS, she rarely describes anything other than expert performance on each of these competencies. In addition, Benner identifies the 'advanced beginner' stage of development as sufficiently competent to enter the nursing profession. The descriptions of competency required to enter the speech pathology profession, as described within current Australian assessment formats, resemble her description of 'competent'.

5.2.3.2. Handling Increasing Complexity

Current workplace performance assessment tools in speech pathology appear to have an implicit understanding that competence is reflected in the ability to handle increasing complexity. For example, ratings of the students' adaptability and creative thinking on the Assessment of Clinical Competence (The University of Sydney) describe a change from:

> Category 0: "Student rigidly follows prescribed procedures; student has no awareness of not meeting client needs and of the need for change".

Through to:

> Category 4: "Student identifies the need for change in client behaviour, conduct of session or of their own behaviour during the session, and makes an appropriate and creative response"

The concept of developing skills that can be applied to increasingly complex situations also appears to underpin the Clinical Eduction Evaluation Form (The University of Queensland). The development of competency for each skill on this tool is reflected by the increasing complexity of tasks or performance students can carry out successfully. This notion of increasing complexity is also evident in the descriptions of increasing competency in Table

8 for generic ratings of independence, collaboration, complexity, efficiency, and critical/creative thinking.

McAllister and Rose's (2000) article describes an approach used to teaching clinical reasoning over a four-year program at Latrobe University which addresses the need to develop the ability to handle increasing degrees of complexity. This teaching program focuses on facilitating development of clinical reasoning skills in an applied fashion, integrating knowledge and practice from other aspects of the curriculum, and features a graded progression in applying clinical reasoning skills to increasingly complex situations. Thus the cases used progress from simple case examples to cases with several parameters requiring integration of knowledge from several theory streams. Then cases are introduced that require attention to be directed to the context in which the client functions and finally scenarios including professional issues such as the workplace environment, ethics, staff issues, legal, and safety requirements.

This notion that a competent professional being able to handle complexity is well described in the literature and, as summarised in the literature review, relates to competency resulting from complex professional judgement informing action in dynamic workplaces. That Australian speech pathologists equate developing competency to the ability to manage more complex tasks is closely related to the commonly accepted understanding in educational literature that learning is reflected by the ability to handle increasingly complex information.

5.2.3.3. Bloom's Taxonomy

As described in the literature review, Bloom's Taxonomy (Bloom, 1994; Carter, 1985; Clark, 1999; Krathwohl, 1994) identifies three domains of learning and a developmental progression of behaviours, from simple to complex, which indicate learning has occurred over these domains. Each domain (cognitive, affective, and psychomotor) has subdivisions, starting from the simplest behaviour to the most complex, and is seen as degrees of difficulty where one must be mastered before the next one can occur (Clark, 1999).

The Bloom taxonomy was originally developed to assist educators in their development of assessment questions that give students opportunity to demonstrate the range and complexity of their learning rather than simple regurgitation of facts (Biggs & Collis, 1982). Bloom's taxonomy has been applied to the assessment of competence in speech pathology in the form of 'Indicators of Emerging Competence' or IECs, which describe beginning, intermediate, and advanced level behaviour descriptors for each of the 1994 version of the CBOS Units and

Elements (McAllister et al., 1996). This analysis does not claim to be exhaustive but rather to provide some examples of behaviours that represent different levels of performance on the CBOS and uses some of the Bloom descriptors to illustrate behaviours at the different levels.

However, the IECs are limited in their analysis of competence as they relate only to the knowledge domain, appear to intermix higher and lower level descriptors, and are strongly influenced by notions of independence. It would appear that applying Bloom's taxonomy to the CBOS units and elements that the task is not as simple as would be hoped. Applying descriptors from all three domains is complex and whether other concepts should be used to influence the application of the Bloom descriptors needs be considered.

5.2.3.4. SOLO Taxonomy

Biggs and Collis (1982) have developed the Structure of Observed Learning Outcomes taxonomy (SOLO) that aims to assess learning outcomes rather than assist the development or design of assessment strategies (as per Bloom's taxonomy). SOLO qualitatively describes (or assesses) student learning in terms of the structural complexity of the learning outcome. The five levels detail a progression over three dimensions. The first dimension is termed 'capacity' and looks at the way in which students relate the question cue to all the potential data available, consider how this information interrelates, and entertain a range of potential hypotheses.

The second dimension is titled 'relating operation' and describes the way in which students can extend their thinking beyond the concrete specifics of a situation. The final dimension is called 'consistency and closure' and describes the way in which students handle inconsistency and the need to 'close' or offer a final decision or answer. Thus the taxonomy analyses how well the response relates to the initial cue in terms of how much and what type of data is included in the students' response and then how well each piece of data is related to another and whether an open-ended response or hypothesis is generated. Biggs and Collis (1982) identify that they have two main effects in mind: knowledge (content) and cognitive processes that are induced by the proper understanding and application of the subject.

A more simplified description of their taxonomy is a cycle of learning that progresses through the following stages (Biggs & Collis, 1982):

1. Prestructural: below the modality of learning in question.
2. Unistructural: one aspect is recognised.
3. Multistructural: several aspects are recognised but not integrated.
4. Relational: the totality is put together.
5. Extended abstract: "...a whole new ball game" (pp. 231).

These concepts would appear to be compatible with the concept of developing competence in handling complexity that appears to underlie the Australian speech pathology profession's conceptualisation of the continuum underlying the progression towards entry level competence. In addition, an assessment proforma SOLO was found to have positive effects on learning as it was associated with high intrinsic motivation and learning strategies involving search for meaning and avoidance of rote learning detail (Biggs and Collis, 1982). SOLO has primarily been used to assess acquisition and application of propositional knowledge, including in the area of speech pathology (Scholten, 2000). However, Biggs and Collis also put a strong case for the application of SOLO to any new learning episode including applied situations. Thus it would appear that the SOLO taxonomy holds some promise as a framework to develop descriptions of clinical performance over the continuum from beginner to competent.

5.2.4. Summary

Review of the literature and current practice has not yielded any clear answers to the two main questions posed at the start of this analysis regarding how the rating scales for the assessment tool should be designed. First, what dimension or aspect of a person's performance on a particular skill should be measured (rated) to determine their competency? Second, what does the progression from a novice to competent or entry-level practitioner in speech-language pathology involve?

While rating scales are used frequently for performance assessment in the health sciences, the literature has few descriptions on what dimension of performance is being rated. Those that do, provide no details on what behaviours represent what levels of performance on these scales. Current Australian assessments often appear to be rating competencies on a dimension that could in fact be described as a competency in itself e.g. efficiency. In general, ratings on particular points of the continuum of beginner to competent seem to be based on a description

of competence in managing increasingly complex situations and cases, with reference implicitly or explicitly to the concept of independence or interdependence. In some cases a description based on a hierarchy of behaviours (rather than increasing complexity) that reflect increasing competence is also described, for example, for the concept of 'interdependent learner'.

Independence or interdependence reflects the degree of supervision, support, or guidance that students require to function competently and is often linked to appropriate seeking of support. This appears to be an influential concept in Australian speech pathology performance assessment. The second concept that appears to strongly influence Australian assessments is the idea that, as students approach entry-level competency, they are able to handle increasing degrees of complexity.

Benner's application of Dreyfus & Dreyfus' work on the development of performance skills in nursing appears to be the only theoretical model representing progression of skill performance from beginner through to competent in health sciences (Benner, 1984). Application of her work to assessment of competence in speech pathology is hampered by the fact that she generally only gives examples of the expert levels of performance for each specific competency she has identified for nurses (Benner, 1984; Benner et al., 1996). Notwithstanding this, the Dreyfus' model (1996) and Benner's work offer some useful guidelines for judgement of increasing competence and a description of what these behaviours may look like. This conceptualising of competence could be usefully integrated with models of how propositional knowledge is organised and retrieved in relation to specific competencies such as clinical reasoning (Boshuizen & Schmidt, 2000; Newble et al., 2000).

Finally the educational literature does offer some useful taxonomies for assessment of learning. The Bloom taxonomy (Clark, 1999) provides the useful and enduring notion that competence is comprised of knowledge, skills, and attitudes and that this develops in complexity. The SOLO taxonomy (Biggs and Collis, 1982) offers a simple paradigm for analysing responses in a given situation for the degree of complexity represented in the learning that has been demonstrated.

In summary, it would appear that there is no agreement on what dimension or aspect of a person's performance on a particular skill should be measured to determine their competency. On the other hand, there does seem to be some hierarchies mapped out that could be applied to illustrate the progression from a novice to competent or entry-level practitioner on the skills

of speech pathology and some strategies to assist with developing descriptors where they have not been already designed.

Thus, in the absence of one clearly superior approach, it was decided that a pragmatic strategy be used that rates students' performance on a skill across a dimension that describes levels of competency and cannot be confused with another competency dimension. Each band or point of the scale would have behaviourally anchored descriptors that outline what performance would be expected from students at the various stages of development (as previously identified in the description of the design phase). The descriptors would be developed with reference to developmental hierarchies where they have been mapped, either through expert consensus or as part of a theoretical model, and use the following processes where new hierarchies of behaviours need to be developed through:

1. Applying the concept of a developing ability to manage complexity, that include knowledge, skills, and attitudinal aspects (Bloom, 1994) and are mapped out using the SOLO taxonomy (Biggs & Collis, 1982) as a framework.
2. Integrating an understanding of how the development of expertise also involves the development of knowledge through experience and transformation in how this knowledge is used in clinical situations to inform judgement, as described by Benner (1984, 1996) and Dreyfus and Dreyfus (1996).
3. Attending to the degree of support/guidance students require to perform a skill competently.

The third area for consideration regarding assessment design was identifying what generic competencies should be included in the assessment.

5.3. Generic Competencies

As outlined in the literature review, generic competencies are conceptualised as arising from combinations of knowledges, skills, and personal qualities and to enable the holistic integration and coordination of occupational competencies into competent professional practice. It was clear from the literature, consultations with the expert group, focus groups, and SPAA conference consultations, that including some generic competencies in the assessment design was strongly supported. The CBOS and current Australian speech pathology assessments of student workplace performance were also examined to identify generic competencies relevant to the practise of speech pathology.

5.3.1. Literature

The literature relevant to professional competencies, particularly those evident in health science practice, referred to many different types of generic competencies both implicitly and explicitly. A broad classification of the competencies identified as core generic competencies in the literature is listed in Table 10. As can be seen from this table, it was not particularly obvious as to how these generic competencies could be classified or organised and applied to the practice of speech pathology and a great deal of overlap and similar terminology is used.

Of most concern to this project however, was that there was no single taxonomy that lent itself to specifying the generic competencies essential to the practice of speech pathology. Fortunately the profession itself had a clear understanding of what these competencies might be as the SPAA Conference consultations reached an undisputed consensus as to what generic competencies should be represented within the assessment.

Table 10. Generic Competencies Represented in the Literature

General grouping	Generic competencies included in grouping
Interpersonal skills E.g. ABIM (1998), Higgs & Titchen (2001), QAAHE (2001), Sharpley (1997).	• Working with others and in teams • Professional relationships • Personal and professional skills
Critical thinking E.g. ACER (2001), Benner et al. (1996), Higgs & Hunt (1999), Higgs, Jones & Refshauge (1999), Hunt et al. (1999), Hunt & Higgs (1999), Johnson & Shewan (1988), McAllister (1997), McAllister & Rose (2000), QAAHE (2001), Tracy et al. (2000).	• Problem solving • Clinical reasoning • Clinical judgment • Analytical skills • Collecting, analysing and organizing information • Self evaluation • Ability to apply theoretical knowledge to practice
Communication skills E. g. ABIM (1998), ACER (1998), Fleming & Mattingly (2000), Sharpley (1997).	• Written communication • Communicating ideas and information • Counselling, interviewing • Communication/interpersonal skills that enable client's perspectives to be involved in decision making
Personal attributes E.g. ABIM (1998), Epstein & Hundert (2002), Higgs & Hunt (1999), Hunt et al. (1999), Hunt & Higgs (1999).	• Orientation to serving and improving society • Being accountable • Recognizing limitations • Tolerance • Integrity • Sensitivity, respect for and empowerment of clients • Awareness of value judgments
Lifelong learning skills E.g. ACER (2001), Higgs & Hunt (1999), Higgs & Titchen (2001), McAllister & Rose (2000), Sefton (2001).	• Self reliance in acquiring knowledge and problem solving • Critical self appraisal and reflection • Metacognition
Professionalism E.g. Higgs & Hunt (1999), Higgs & Titchen (2001), Loomis, 1985b, McAllister & rose (2000). Moral and ethical behaviour E.g. Epstein & Hundert (2002).	• General professional conduct • Professional judgment • Ability to interact with and change the context of practice • Professionally responsible
Managing the workplace E.g. ACER (2001), Luttrell et al. (1999), QAAHE (2001), Sharpley (1997).	• Planning and organizing activities • Leadership

5.3.2. Generic Competencies Identified at SPAA Conference Consultation

A remarkable degree of consensus was reached independently by 4 groups of speech pathologists participating in the consultations at the SPAA conference. The priority generic competencies identified were detailed in Chapter Four (Section 4.2.3.3.) and lent themselves to classification under the following 4 headings with the remaining two competencies lending themselves to inclusion under other generic or occupational competencies. An extended listing of the suggestions made at this consultation is included in Appendix 13, Table 11 provides a summary.

Table 11. Generic Competencies Identified at SPAA Conference Consultation

Generic Competency	Sub-competencies
Clinical Reasoning	• Judgement/decision-making/problem solving • Moving from theory to practise • Ethical reasoning • Reflection (links to lifelong learning skills) • Intercultural competence
Lifelong learning	• Self-evaluation, reflection, change and learning • Reflective practitioner including: lifelong learning, self evaluation, self praise
Professional Role	• Handling contextual issues • Subsumes – organisational skills, ethics • Personal/self management
Professional Communication	• Communication skills • Interaction/interpersonal skills

Both the CBOS and focus group data were examined to confirm whether this four way classification accurately represented the broader constituency of speech pathology practice within Australia and could be applied equally well to the generic competencies represented by these two sources.

5.3.3. Generic Competencies Represented in the CBOS

Initial analysis of recurring key words, phrases, or concepts that were represented in the introduction to CBOS and at the performance indicator and cue levels of the CBOS suggested that the profession saw itself as having competencies that were based on more than the competent performance of specific skills. A number of broad themes were apparent within

CBOS and given that the CBOS represents the professions' view of what it believes it does, and can be assumed to represent what the profession values about itself, these are likely to have strong validity for CEs and speech pathology students. These themes were represented across a variety of the occupational competencies detailed by CBOS suggesting that these generic competencies are conceptualised as enabling appropriate performance of the occupational competencies.

The generic themes were initially broadly categorised as representing attitudes, specific skills, or activities in relation to research and promotion of the profession and generic competencies such as lifelong learning, ethics, communication, clinical reasoning, teamwork, professional competencies, a holistic orientation, and generic abilities related to maintaining appropriate levels of skill performance. However, the four categories identified at the conference consultation provided an excellent classification framework. A number of the themes could be represented under several headings, which was not surprising given how interrelated these competencies can be expected to be, and so were placed in the category that seemed most salient although this was subject to minor change later when developing the generic competencies in detail. Thus the competency of 'clinical reasoning' was illustrated by statements requiring that an entry level speech pathologist should be able to give rationales, demonstrate clinical reasoning, or critically evaluate the literature. Professional competencies were most frequently mentioned and could be classified under themes related to attitude, ethics, teamwork, behaviour, holistic orientation, and skill performance. This classification is summarised in detail in Appendix 14.

5.3.3.1. Generic Competencies Identified Through Focus Group Consultations

A number of comments and suggestions from the focus groups were clearly related to generic rather than occupational competencies. The questions that generated this type of response were usually the following:

1. What are the most important features of a student's performance that help you decide whether they are performing competently, well or poorly?
2. What do you think is the most difficult aspect of a student's performance to assess?
3. What aspects of a student's performance do you think are most critical to assess?

Other responses to these questions that were not classified as generic competencies related to more specific skills covered under the CBOS e.g. knowing the general principles for giving

formal tests, reporting skills. Some responses described dimensions of skills previously discussed as relevant to tracking the development of competence such as managing complexity, automaticity, or focussing on the client and not themselves. Once again, the four generic competencies identified at the SPAA conference were applied as a framework for classifying data from the focus groups that could be described as identifying generic competencies (knowledges, skills, and personal qualities that combine to enable students to apply or develop occupational competencies). The framework from the conference consultation provided an excellent classification of all relevant data from the focus groups.

Table 12. Classification of Focus Group Themes According to Four Generic Competencies

Competency	Related focus group themes
Clinical Reasoning	• Clinical problem solving • Applying theory to practice • Critical thinking • Contingency planning, ability to change and monitor within situations • Integration of theory to practice • Flexible • Creativity (Including statements related to lateral thinking, changing within sessions, ideas, session materials varied and interesting, adapting resources, adapting to client needs, change track and still achieve goals, contingency planning • Ability to see the whole, to synthesise
Lifelong Learning	• Initiative e.g. willing to learn, goals and plans, get involved, asks questions, offer ideas, general independence in work setting • Independence in finding information • Recognising what they don't know and have to find out • Self directed learning • Evaluate performance against own goals • Good self assessment and self evaluation • Integration of new information and feedback • Ability to analyse • Demonstrated ability to learn, ability to change, acquire skills • Transfer of skills • Knowledge of limitations

Professional Role	• Organisation, Time management • Professional responsibility/accountability • Professional behaviour e.g. demeanour around department, attitude, balancing knowledge of limitations with confidence • Commitment • Attitude • Taking responsibility for clients • Efficiency/time management skills • Holistic • Focus on client outcomes/client care
Professional Communication	• Communication skills i.e. as an overarching skill with clients and colleagues, should be rated against all items • Ability to communicate within a team, work within a team • Interpersonal skills • Rapport with clients and staff • Interpersonal skills with client and workplace

5.3.4. Generic Competencies Represented in Current Australian Assessments

It was noted earlier that a number of dimensions in current Australian assessments of speech pathology students' workplace competencies ask CEs to rate the students' development on performance dimensions which could themselves be described as generic competencies (Table 8). These include: interdependence; professionalism; adaptability and creative thinking; self-evaluation; collaboration; and critical/creative thinking. The remaining dimensions of independence and the closely related dimensions of complexity and efficiency have already been identified as useful concepts to describe progression towards entry-level competency, rather than competencies in themselves.

The rating dimensions that are identified in Table 8 as closely related to competencies can be classified into the 3 out of the 4 generic competency headings as follows. Critical or creative thinking can be classified as clinical reasoning competencies, self evaluation and being an interdependent learner relate to lifelong learning competencies, and collaboration and professionalism are professional role competencies.

Interestingly, the fourth category of professional communication competencies is well represented by those assessments that are organised such that they rate students on more general competencies, where CBOS is implied. For example, the Clinical Education Evaluation Form (The University of Queensland) has a section title "Communication Skills",

The Clinical Protocols (Flinders University) has a unit titled “Interpersonal and Counselling Skills” and the Clinical Assessment Form (Latrobe University) has a section titled “Interpersonal Skills”. These assessments also include other competencies that could be categorised as clinical reasoning (problem solving, planning, interpretation, diagnostic process/making recommendations), lifelong learning competencies (student centred learning, self-evaluation, professional development), and professional role competencies (organisation and professional responsibility, professional development, professional skills).

5.3.5. Final Classification of Generic Competencies

The following classification appears to effectively account for the generic competencies identified from multiple sources of evidence including focus groups, current Australian assessments and key words, themes, and phrases in CBOS and the literature:

1. Clinical Reasoning.
2. Lifelong Learning.
3. Professional Role.
4. Professional Communication.

5.4. Attending to the Impact on Learning

As identified previously in Section 3.1.2. of the literature review, it is clear that the assessment process directs students’ attention to their learning and thus assessment design can impact upon this learning positively or negatively. This concern for students’ learning was also identified by the focus groups identifying the need for formative and summative assessment processes:

> *What I would like to see is actually maybe have one where you could include a formative as well as a summative component, so rather than filling in two feedback forms at different spots, you can actually see where they are at mid placement and at the end.* (CE, LaTrobe focus group)
>
> *A ‘mini-mid’ assessment tool that would look at learning.* (CE, The University of Newcastle focus group [field note])
>
> *...well pretty much all the assessment comes right at the end, so you get a tiny little bit of assessing along the way ... But it tends to all come in a big lump at the end. And I think there is a lot of room for doing that, carrying out that more continuous assessment along your placement.* (Student, Flinders paired interview)

It's moving away from it being achievement and being focussed on the learning. (CE, Adelaide focus group)

The assessment needs to be designed with this in mind. First, a more detailed formative component should be included as discussed in the literature review (Section 3.1.2. of Chapter 3). It is proposed that this would require ratings at both the unit and element level of the CBOS and Generic Competencies plus an overall rating and be carried out mid way through the placement. More detail should be included in the formative assessment to ensure that the students are given an opportunity to participate in thoroughly assessing their current level of competency and planning for their learning over the remainder of the workplace practicum. The summative component of the assessment will be designed to be a briefer summary of what the students have achieved over the placement subsequent to the formative assessment.

Second, as described, considerable attention should be paid to the content of the assessment to ensure it clearly reflects the competencies required and describes the progression through these competencies such that the students' attention is directed appropriately and matches the perception of students and CEs of the task at hand. This is particularly important with regard to the Generic Competencies (described in Section 5.3 above) and the identification of behaviours that illustrate levels of competence (see Sections 5.1.2. and 5.2.3. above, and 5.7.2 below) as these are the aspects of the assessment tool that will be newly developed as part of this research and will be unfamiliar to CEs and students, unlike the CBOS. Explicit description of these aspects of the assessment tool ensuring this content is communicated clearly will allow the students and CEs to identify whether these aspects of the assessment tool do in fact match the values and attitudes they hold and consider to be important to the profession.

Third, an opportunity to make comments should be provided for each unit of competency, both at mid and end placement assessments. This is suggested on the basis that, during the focus groups, all students strongly indicated that general comments from their CEs were highly valued and many CEs indicated that they found making qualitative comments an important part of the assessment process.

5.5. Supporting Judgement

The fifth and final component of assessment design considered was the need to ensure that the CEs' judgement of the students' competency was well supported. Training of raters is the most obvious strategy for supporting this judgement and appropriate rater training for

performance appraisal has been found to positively influence the accuracy of rater's judgements (Woehr & Huffcutt, 1994). On the other hand, given that current practice in Australia means that students are frequently placed with field educators who are unpaid volunteers and provide placements in addition to their current workload and the difficulty of accessing training for those working across vast distances in rural areas, it is neither feasible or reasonable to compel the CEs to attend training sessions on assessing students.

Overall it was decided that the tool ideally needed to demonstrate good validity regardless of the level of training that raters may or may not have been able to access. This meant that it was particularly important that the tool design supported CEs' judgment effectively. The two scenarios were identified by the researcher and expert group as situations where CEs potentially may require particular support included novice CEs who lack of experience in making judgements as well as making judgements about students whose performance is of concern or marginal.

With regard to the first scenario, the evidence regarding whether inexperienced CEs do have more difficulty in making judgements of student performance than more experienced CEs is mixed. Intuitively, given that making judgements is proposed to be a developmental skill that is informed by experience (Down & Hager, 1999; Dreyfus & Dreyfus, 1996), it seems reasonable to assume that more experienced CEs will be better able make judgements of competence. Friedman & Mennin (1991) suggest that an examiner's frame of reference is established through experience with examinees. Jones (2001b) asserts that each assessor's understanding of the standard or quality of performance is formed through experience and that they require considerable expertise in their own field. Chapman (1998), writing about assessment in speech pathology, also suggests that (particularly if there are no clear criteria against which to judge) the assessor makes comparative judgements of performance based on previous experiences of student performances. In this situation, the less experience CEs have, the more idiosyncratic that experience is likely to be.

However, Cross et al. (2001) actually found that experience did not affect physiotherapists' ability to assess student performance on video vignettes, assessors whose judgement closely matched that of university clinical education staff were not differentiated according to experience with assessment. Conversely Landy & Farr (1980) found that increased rater experience positively affects quality of performance ratings and that raters who are judged as better performers in their jobs are better at rating the job performance of others.

Nevertheless, lack of confidence in making judgements about competent performance is frequently reported in the literature (Chapman, 1998) and it seems likely that novice CEs, particularly those with little professional experience as well, will feel even more anxious regarding their ability to make a judgement of competence. This issue is reflected in McAllister and Lincoln's (2004) developmental hierarchy describing the progression from novice to expert based on their experience in working with CEs. The other components in this hierarchy reflect many of the aspects involved in the development of expertise identified for the Behavioural Descriptors for the assessment tool.

Anxiety around judgement can be extreme with regard to students who are having difficulty developing competency in their performance of workplace skills and further complicated by affective reactions to failing students (Ilott & Murphy, 1997). Failing or marginal students are a low incident event of high salience. The implication for assessment tools is that they need to be efficient for use with 90% of students and highly effective for the 10% about whom concerns are held (Hunt, 1992). This view was echoed by CEs in the focus groups who commented that they wanted the assessment tool to be both brief and detailed – while acknowledging that one prevents the other.

> *If it is too long it doesn't get used appropriately. People then can't be bothered reading all the information.*
> *If it is too short you leave things out, things don't get covered so people feel like they haven't had enough chance to put forward their ideas.*
> *Let's say in the middle.* (Conversation among CEs Latrobe focus group)

This led to the decision that ideally the assessment should consist of a brief, easily understood format supported by layers of detail and information that can be accessed when making more specific and complex decisions about students' competency. Given the potential advantages of computer technology for meeting this need and the suggestion by some CEs and students in the focus groups that a computerised version of the assessment tool would be useful, it was decided to also consider the possibility of a computer based assessment tool.

Ilott & Murphy (1997) suggest that making a judgement of 'fail' for a student depends on clear definitions of threshold standards or competence. This was confirmed by focus group comments, for example:

> *I think it goes back to what H was saying much earlier to have very clear guidelines as to what is the pass criteria, even for things that are maybe judgement calls, that there*

> *are actually specifics on what sort of actions will constitute that judgement being that, or made.* (CE, Latrobe focus group)

There is clearly a great deal of merit in this observation and this can certainly be addressed in the way that entry level competence is described to inform the final assessment of students prior to graduation. Identifying students who are struggling in their development of competence earlier in the course is clearly preferable for reasons of counselling and remediation, and ideally in the future the assessment tool can be benchmarked by individual university programs to match their particular theoretical and practical curriculum. Ilott & Murphy (1997) and Duke (1996) also indicate that CEs are generally able to identify marginal students, suggesting that there should be an opportunity for them to indicate this on the assessment format.

Review of the literature indicated that marginal students tended to show global issues in their work, often related to Generic Competencies. These were described as difficulties with lifelong learning e.g. responding appropriately to constructive feedback and changing behaviour, professional skills such as being prepared and organised, poor communication and interpersonal skills, and inadequate clinical reasoning skills (Maloney et al., 1997; McAllister & Rose, 2000). This viewpoint was evident in comments by CEs in the focus groups about identifying students who are marginal:

> *Integration. The failing students just couldn't integrate, they can't integrate theory into practice and they can't transfer from one client to another, they can't transfer skills, they can't generalise.* (CE, Adelaide focus group)
> *There wasn't anything that really stood out, it was like a whole package of things. ... There wasn't just one thing I could put my finger on. It was a global thing really.* (CE, Adelaide focus group)

This further supports the inclusion of generic competencies along with the occupational (CBOS) competencies in the assessment.

Overall it is suggested that judgement needs to be supported by assessment formats that provide a rich source of information and context to guide the assessor's judgement (Jones, 2000, 2001b). Development of scoring rubrics that focus on observed qualities of performance can improve the psychometric qualities of an assessment (Wolfe & Gitomer, 2001). CEs also tend to express concern regarding assessment of areas that cannot be easily objectified such as attitudes, values, and caring as opposed to practical skills (Duke, 1996), so ensuring that these are well described in the documentation may also be helpful. At the same

time it is important to leave room for the exercise of professional judgement and not over specify the competencies or outcomes of interest by resorting to checklists (Jones, 2000, 2001a).

In addition, some writers suggest that gathering data from self, peer, patient, and colleague assessments in the workplace are potentially a source of useful information (Dauphinee, 1995; Higgs et al., 1999; NMBE, 2002). These types of assessment would provide further data to assist the CE's judgement and probably increase the construct representativeness of the assessment overall.

Research consulted regarding the validity of such assessments was mixed. Peer ratings were used to assess physician performance as part of ongoing registration to practice and were found to be meaningful and well received by the assessees (Ramsey et al., 1993). Ratings of medical knowledge by peers were also highly related with the assessees' examination scores although ratings of humanistic qualities were poorly related with exam scores. Davis (2002) found that while there were generally good correlations between the university staff and peer ratings of obstetric and gynaecology residents, there were poor correlations between these ratings and ratings by the resident themselves and nursing staff. Finally, ratings from standardised patients used in OSCE are often included in judgements of student performance (Dauphinee, 1995) but no information was found as to how these compared to other assessments made during these examinations.

Given that the literature suggested including self, peer, colleague, and patient assessments in the determination of competency, this concept was raised during focus group discussions – and unanimously greeted with concerns regarding its validity and practicality, although it was acknowledged that it had formative value. Given these concerns and the lack of clear direction from research it was decided that incorporation of assessments from other sources should not be pursued for this assessment tool.

Overall it is important to ensure that the assessment is based on evidence of sufficient quantity and quality through formative assessment and ongoing feedback and consideration of performance over the workplace placement (Peters et al., 2001). This should include observing students working with as wide a variety of cases as possible to avoid the pitfall of case specificity that has been consistently identified as having a significant negative impact upon the validity of OSCE assessments. Case specificity refers to the fact that students' performance on one case does not predict their performance on another – suggesting that competence, at least in clinical reasoning, is not necessarily transferable from one situation to

another (Dauphinee, 1995; Newble et al., 2000). This supports the intention to design the assessment to be used by CEs who are responsible for students in the workplace and thus have the most opportunity to observe and judge the students' competence through frequent and regular contact with students to ensure their learning needs are met and clients' wellbeing is safeguarded. In addition, CEs should be able to indicate if a competency is not observed.

5.5.1. Summary

Reviewing sources of evidence relevant to supporting CEs' judgement of students' performance suggested that strategies related to other aspects of the research design would also effectively inform CEs' judgement. These aspects include careful attention to wording throughout the assessment to ensure clarity, behavioural descriptors that focus on qualities of performance and not specific behaviours, inclusion of generic competencies, a formative and summative component, as well as the assessment being conducted by the CE who has the maximum amount of evidence on which to base a judgement. In addition a brief, easily understood format is required, that is further supported by layers of detail and information that can be accessed if needed, and the notion of a computer based assessment tool merited consideration. An opportunity to indicate if students' performance might be marginal is also required as well as if a competency has not been observed.

5.6. Summary of Assessment Design Parameters

The multiple and reiterative processes of investigating various sources of evidence resulted in the decision to incorporate the following aspects into the research assessment format:

1. Global ratings of qualities of performance will be used rather than checklists of specific types of task behaviours.
2. Clear descriptions and examples will be provided as to how these qualities may be evidenced in behaviour.
3. Rating tasks will not require multiple complex specific judgements.
4. The same rating process will apply to each competency.
5. Ratings will be used that represent a developmental progression towards entry-level competency.
6. Behavioural descriptors will be used to guide these ratings and will include:

    a. Managing complexity, drawing on concepts from Bloom's Taxonomy (Bloom, 1994; Clark, 1999) and the SOLO taxonomy (Biggs & Collis, 1982);
    b. How competence is transformed over time through the integration of knowledge and experience, drawing on concepts from Benner (Benner, 1984) and Dreyfus & Dreyfus (Dreyfus & Dreyfus, 1996);
    c. Attending to the degree of support or guidance students require to perform a skill competently.
7. A VAS will be used that has 3 anchors of novice, intermediate and entry-level. The intermediate anchor would be placed in the centre of the line without a specific mark to nominate a point on the line.
8. Checkboxes will be included to indicate:
    a. Above entry level performance;
    b. If a competency has not been observed;
    c. If a performance may be marginal.
9. The assessment will include a detailed formative and less detailed summative assessment.
10. Generic competencies will be included along with CBOS competencies and will be developed with reference to all sources of data.
11. The content of the assessment will reflect the perceptions of CEs and students as to what is important to the development of competency, as well as information derived from current research and expert opinion.
12. Opportunities to make comments will be provided for each unit of competency for both formative and summative assessments.
13. The assessment format will include layers of detail to support CE's judgement, should that be required.
14. The development of a computer based version of the assessment tool be considered.

5.7. Final Assessment Tool Design

The final format for the assessment tool consisted of an assessment booklet (Appendix 15), an assessment resource booklet (Appendix 16), and an alternative online assessment (which will be described in more detail in Chapter 6). Implementing the design parameters required selection of a physical format for the assessment booklet, finalising the behavioural descriptors, specifying the Generic Competencies as well as development of materials for the resource booklet, as well as an online version of the assessment tool.

5.7.1. Page Layout

The page layout of the APTA tool format (Roach et al., 2002) was adapted and used for the research assessment tool as the APTA assessment had been extensively researched and empirically validated and was also closely aligned with the design requirements of the assessment under development. Demographic sheets were included for research purposes (as will be described in the methodology of Phase 2) and a brief explanation of how to use the tool. Resource materials were included in a separate booklet, unlike the APTA tool that includes resource materials in the assessment tool itself.

5.7.2. Behavioural Descriptors

As described previously, it was determined that the students' performance would be rated on a dimension that represented levels of competency, the same rating decision would be applied to each competency, and would not involve multiple complex judgements. These levels of performance were described behaviourally and related to the three anchor points on the VAS scale (novice, intermediate, and entry-level) and termed "Behavioural Descriptors". The Behavioural Descriptors were developed with reference to theories informing the ability to manage increasing complexity; the transformation of understanding a situation and making judgements that occur with increasing knowledge, experience, and the degree of support/guidance required to perform competently.

The behavioural descriptors were refined and condensed through circulation for comment by the expert reference group and were placed on the page opposite the VAS so that they were constantly present to refer to while rating. A more detailed version of the Behavioural Descriptors was also included in the resource materials.

5.7.3. Specifying the Generic Competencies

Clearly to assess students on the four generic competencies, it was necessary to specify what was included within each category and describe these in terms that would assist in identifying behaviours relevant to each competency. For consistency, it was decided detail each competency using the format of the CBOS, as follows:

> **The Units** are broad areas of professional activity. They are not sequentially ordered and do not imply any stages or isolated steps in the process of practice. The practice of

> the profession is multidimensional and the numbering of the Units is for reference only.
>
> **The Elements** are more specific activities carried out within the unit**.**
>
> **Performance Criteria** have been developed in order to be able to infer whether the elements of competency are being carried out to an acceptable standard.
>
> **Cues** illustrate the knowledge base, practical considerations, actions, attitudes, and some contextual features that are required as evidence that performance criteria have been achieved.
>
> (pp. 4) (SPAA, 2001)

The sources already described were used to develop and organise the detail of each of the four generic competencies. Literature related to specific competencies was consulted in addition to the literature outlining generic competencies as already described. This included literature in relation to clinical reasoning (ACER, 2001; Ferguson et al., 2001; Henley & Twible, 2000; Higgs & Jones, 2000; Higgs et al., 1999; Mattingly, 1991; Pithers, 2000; Refshauge & Higgs, 2000), lifelong learning (Candy & Worral-Carter, 1999; Ferguson & Fitzpatrick-Barr, 2001), professional skills(ATEAM, 2001; Lincoln, 2002; SPAA, 2002), as well as information related to all of the generic competencies including professional communication skills (Benner et al., 1996; QAAHE, 2001; Twible & Henley, 2001).

Once the researcher had drafted a detailed breakdown of each generic competency, the draft was circulated for comment to the expert group as well as a number of university based CEs or lecturers responsible for the clinical practicum or teaching of professional skills. A number of minor changes to wording were identified and made, with no changes required to the next draft that was ultimately included in the research assessment format.

5.7.4. Resource Material

The final version of the resource material was collated into a booklet separate from the assessment booklet and titled “Assessment Resource Manual”. It included a more detailed version of the Behavioural Descriptors, developed during the process of refining these descriptors, and a copy of the Generic Competencies detailed according to the CBOS format.

In addition, the Behavioural Descriptors were applied to each unit of competency to provide a description of what kinds of behaviours may illustrate a novice, intermediate, and entry-level quality performance. This analysis was circulated to the expert reference group and several CEs considered to be expert in their field. A number of minor modifications to wording were required before inclusion in the resource materials.

5.7.5. Development of the Online Version

Developing a computer based version of the assessment tool that was investigated initially because CEs and students expressed some interest in this option during the focus groups and it seemed likely that it would streamline the process for CEs using the assessment tool during the research. In addition, a computer version offered the possibility of supporting the research process through automatic measuring of the VAS.

Consultations with Portal Australia Inc., a software development company specialising in database development, identified that a computer version had a number of potential advantages. First it would make the assessment process easier for CEs as resources to assist judgement could be available immediately, rather than searching through the Assessment Resource Manual. The computerised assessment would be easier to navigate as it would direct CEs as to what needed to occur next and would dispense with flicking back and forth across pages of the Assessment & Research Materials booklet. Second, if the assessment was provided online, there were many advantages for the research process including all data from the VAS being measured automatically and returned to the researcher, minimising the loss of data through CEs forgetting to post back assessment booklets or booklets being lost in the post. This in effect was also a more secure method of data return. In addition, the process of data collection and the measurement data could be downloaded into an organised data base that could be immediately linked with other databases e.g. demographic data, and the assessment process could be made very interactive. Finally, it was anticipated that a computer based/online version of the assessment tool would be welcomed by university programs and evaluating this as part of the research trial was decided by the research team to be a useful process.

Subsequently, Portal Australia Inc. were contracted to develop an online database in consultation with the researcher and based on the hard copy version of the assessment tool and resources. Portal Australia Inc. recommended using an online or web based system as this meant that CEs would not have to install software or open a program on their own machines,

and would not have to physically copy data onto a disc and post the data back to the researcher. In addition, it allowed for all data to be collated and downloaded and exported into one database as a whole, rather than copying and pasting individual data into a database, thus safeguarding the integrity of the electronic data. The online database also allowed for the researcher to monitor the research process and troubleshoot any difficulties, in association with Portal Australia Inc, as they arose. The main disadvantage was the requirement that CEs have access to the internet at their workplace. The format of the online system is described in more detail in the section describing the research methodology in the following chapter (Chapter 6, Section 6.4.2.2.).

5.8. Summary

A rigorous development process was undertaken to maximise the validity of the research assessment tool through careful consideration of the competencies to be assessed, description of the development of competency, and the appropriate strategies for carrying out this assessment. The next phase of the research involved an extensive field trial to evaluate the validity of the research tool and the methodology involved in the field trial is described in the next chapter.

PHASE 2: FIELD TESTING

CHAPTER SIX

6. METHODOLOGY

6.1. Overview

The aim of this research is to develop a valid assessment tool to measure the workplace performance of student speech pathologists. Phase 1 described the development of the assessment tool, that ensured that aspects that promoted the content and substantive components of its validity were addressed (Messick, 1989, 1994, 1996). The final assessment tool consisted of a rating format, either in hard copy or online, and a resource manual. The rating format addressed competencies considered integral to the appropriate practice of speech pathology at entry level and required ratings according to behavioural descriptors that described the development of this competence. The resource manual contained information to inform and support the CEs' judgement of the behavioural expression of these competencies by students.

The primary research question was therefore:

> Does the assessment tool developed over Phase 1 of this research provide a valid assessment of the workplace performance of student speech pathologists?

Phase 2 of this research aimed to answer this question through field testing the assessment tool with a wide variety of students and their CEs across Australia. This field trial was designed to provide both qualitative and numerical data for evaluation and, where appropriate, statistical analysis. This evaluation and analysis aimed to evaluate four of Messick's (Messick, 1989, 1994, 1996) six validity criteria: substantive; structural; generalisability; and external validity.

The methodology used to field test the instrument is outlined in this chapter and the analysis of the data derived from the field testing is described in Chapters Seven and Eight. This chapter begins with a discussion regarding appropriate statistical methodologies for validating assessment tools. The process of determining the statistical analysis refined and focussed the research questions required to elaborate the primary research question outlined above. These research questions and hypotheses are described as well as the process undertaken for the field testing these. The final section of this chapter will outline the nature

of the sample on which the tool was tested. Chapters Seven and Eight detail the analysis procedures and results gained from the field testing.

6.2. Analysis Method

The validity of an assessment tool is inferred through careful scrutiny of all the factors that contribute to its validity. As outlined in the literature review, Messick's model of interrelated aspects of validity (1989; 1994; 1996) provides useful guidelines to evaluate the validity of performance based assessments and is the one adopted by this research. It was clear from the outset of the project that validation of the assessment tool could only occur through field testing. Thus field testing was planned that would include collecting data from the tool and relevant demographic data to both describe the sample and to examine the relationships between the rating scale data and other variables to assess the tool's validity. The following section outlines the analysis options considered and how they applied to the research.

6.3. Options for Statistically Examining Test Validity

Two main types of measurement models and related statistical procedures were identified that can be used to validate assessment tools: Classical Test Theory (CTT) and Item Response Theory (IRT). While both approaches aim to assess a latent variable and relate it to performance on the test (Embretson, 1999), there are some fundamental differences in their approach and utility for evaluating the validity of a performance assessment tool.

These differences were examined to determine the analysis procedure that would be most effective in evaluating the validity of the assessment tool. Prior to the field testing it became apparent that the statistical analysis to be undertaken needed revision and reframing. The following sections will describe CTT and IRT approaches considered by the researcher and then outline the analysis plan for the field trial data.

6.3.1. Classical Test Theory and Item Response Theory

CTT statistically examines relationships within the sample of raw data generated through observations to determine how well the assessment quantifies the latent variable of interest. The independent variable is assumed to be the amount of the latent variable that the person possesses (or their true score) plus any error occurring during the testing occasion. The true score plus error is assumed to be represented by the subjects' observed scores on the items

and are able to be combined additively to predict the dependent variable or the total test score of each person (Embretson & Reise, 2000).

The IRT family of statistical analysis takes a model-based approach where the data is compared to an ideal measurement model to see if it behaves in the way that the model predicts. If it does not, statistical information is provided to guide the revision of the measurement instrument so that it functions more effectively as a measurement tool i.e. generates data that behaves in the way predicted by the model. The IRT models are further divided into two main groups. Rasch based modelling, a one parameter IRT model, allows for some degree of chance variation affecting the data collected and is particularly relevant to analysis of data from rating scales. The other group of IRT models allow for multiple parameters (and does not apply to this research) and aim for the classical ideal of no error or unexplained randomness in the data generated by the assessment tool (Linacre, 1999b).

There are a number of very strong arguments in support of using Rasch based modelling to guide the development of assessment instruments. However, this approach has received attention only recently in the fields of psychology and social science. The usefulness of the Rasch model was first identified and used by researchers in education and to some degree in health. Its utility is now being recognised in the field of psychology and it has been used in recently developed intelligence tests such as the Stanford-Binet V (Embretson, 1999; Embretson & Reise, 2000).

The primary argument that supports the use of Rasch modelling when attempting to quantify latent traits of human beings is the need to adhere to the principles of fundamental measurement. Michell (1997) argues that physical sciences concerned with measurement and quantification of physical phenomena, such as temperature and velocity, have addressed two fundamental measurement issues: first, proving that the attribute being measured is quantitative; and second, constructing procedures for numerically estimating the magnitudes of the attribute being measured. He suggests that psychology, which concerns itself with the measurement of non physical phenomena, has ignored these basic requirements. Generally, measurement in the non physical sciences has involved assignment of numbers to observed phenomena according to a rule and have been criticised for considering this to be wholly sufficient for accurate measurement of underlying traits in people (Michell, 1997).

This misconception regarding measurement has resulted in test developers persisting in treating the numbers assigned from a rating scale as if they were interval or ratio data without first identifying whether the rules of interval or ratio data apply. Frequently this is seen when

raw scores from scales are summed without verifying that they do in fact possess additive qualities and the total score is treated as if it were a measure (Bond & Fox, 2001). In addition, there are numerous examples of measurements from VAS being subjected to parametric statistical analysis on the assumption that it is interval data comprising 100 equally spaced units of measurement and that the latent variable in question can be quantified to this degree (Linacre, 1998b).

These strategies are confounded further by the fact that the statistics used in CTT are simply descriptions of the raw data, not measures (Bond & Fox, 2001), and some are not entirely equal to the task for which they are used (Clark & Watson, 1995). For example, the coefficient alpha e.g. Cronbach Alpha or K-R 20 are frequently used to establish the unidimensionality of a scale even though they are measures of internal consistency and so have limited utility for this task. They are also imperfect indicators of internal consistency because they are strongly influenced by the number of items used by the assessment tool (Clark & Watson, 1995).

Rating scales based on Likert formats, which are clearly categorical data, are also frequently and inappropriately subjected to parametric statistics and the raw scores summed (Zhu, 1996). Even if nonparametric statistics are used it is usually assumed that the correct number of categories quantifying the data has been identified by the test developer, that the respondents will use them in the manner intended by the developer (Wright, 1999), and the steps between categories are of equal size for every item. Also CTT is unable to take into account item difficulty when calculating test scores (Embretson (Whitely), 1983). Rasch modelling, on the other hand, is able to identify how many rating categories do actually function in the data as a whole and to quantify the size of the steps between the rating categories. In addition it is able to identify on which items it is more difficult or harder than others to get an equally high rating and to quantify by how much.

Moreover, CTT approaches are unable to ensure that the attribute of interest can be quantified and that the assessment tool is accurately measuring the quantities of the attribute of interest. On the other hand, the Rasch model of measurement is based on the assumption that, if the assessment items are effectively sampling a unidimensional trait that is quantifiable, the data will satisfactorily meet the requirements of the model. An algorithm is then applied that enables the transformation of the raw scores representing the observations made into an interval measure (with a stated degree of error) of the quantity of the latent variable possessed by each person assessed. Then this measure can be validly examined via

classical statistical methods to identify whether any probable relationships exist between the quantified amount of competency and other variables of relevance.

A further limitation of CTT is that it results in measures that are confounded with the sample of respondents as the difficulty of the item is defined as the proportion of respondents passing the item (Bond & Fox, 2001). Thus the item difficulty will depend on the ability of the sample on which it is being used and it is assumed that the sample used to validate the test sufficiently resembles the group for which the test is intended (Barnard, 1999; Embretson, 1999). While test development involves attending to this issue through careful sample selection, this process is by no means error proof. Rasch modelling avoids this source of error through estimation procedures that estimate the difficulty of the item independently of the ability of the persons assessed.

CTT relies heavily on the availability of parallel measurements to compare with the results generated by the new assessment tool (Barnard, 1999). This is a particular issue for this research as it was initiated because no valid measures currently exist for these competencies. The availability of parallel measurements is not as critical when using the Rasch model as the process of analysis involves comparing the data to the predicted pattern of data that would be collected if the assessment tool approximates the Rasch model of measurement. In addition, CTT is not able to predict what an individual might do when answering an item, limiting the ability to validate the items (Barnard, 1999).

Finally, CTT assumes that errors are normally and uniformly distributed in persons, have an expected value of zero, and are uncorrelated with other variables (Embretson, 1999). Rasch modelling assumes a certain amount of error will be present in the data due to the inherent variability of human error and provides statistics that identify how much error is present in the data. Thus information is provided explicitly as to how much confidence we can have in the assessment tool as a device to measure the amount of the latent variable each person possesses (Bond & Fox, 2001).

6.3.1.1. Summary

Rasch measurement has a number of advantages over CTT when addressing issues such as substantive, structural, generalisability, and external validity (Messick, 1989, 1994, 1996). The most critical advantage of Rasch measurement is that it enables the abstraction of equal units of measurement from the raw data of observations i.e. scores on the items of an assessment tool. These can be calibrated and then used with confidence to measure human

attributes such as competence in speech pathology practice (Bond & Fox, 2001). Not only are these measures interval in nature, allowing valid statistical examination between the variable of competence and other variables of interest, but measures of item difficulty are separate from the ability of the persons as derived from the ratings on the assessment tool by observers and vice versa.

Thus, structural validity, or the fidelity of the scoring structure to the construct domain being assessed can be evaluated. Once that scoring structure is validated, Rasch measurement provides empirical evidence as to whether the theoretical rationale has enabled the creation of items that identify a unidimensional trait that is in fact assessable (substantive validity). The functioning of the assessment items with different groups within the sample (Differential Item Functioning) can be examined and, if an assessment tool does produce data that meets the requirement of the Rasch measurement model, it can also be assumed that the tool can be used with confidence with different samples thus supporting the generalisability of an assessment. Finally, the need for seeking comparative measures (which currently do not exist for this research) to evaluate the external validity of the assessment tool is minimised, as the Rasch analysis process compares the data generated with an ideal measurement model.

In summary, as pointed out by Wright (1999), Rasch analysis allows test developers to move from ambiguous, raw observations to well-defined, abstract linear measures with realistic estimates of precision and explicit quality of control.

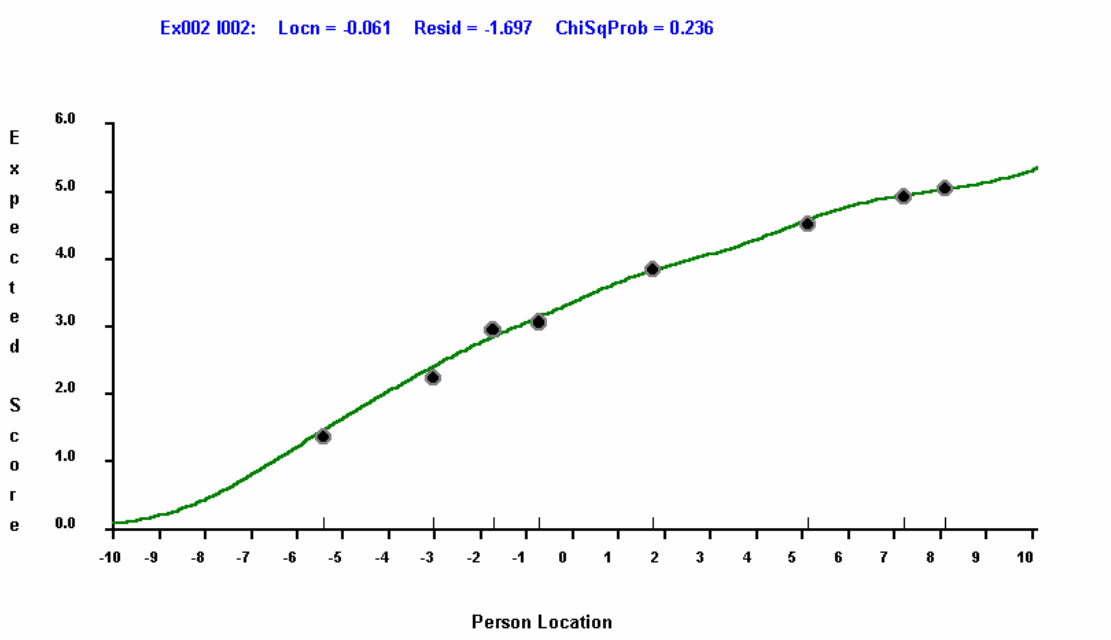
> "Rasch models are the only laws of quantification that define objective measurement, determine what is measurable, decide which data are useful, and expose which data are not." (pp. 80, Wright, 1999)

6.3.2. Description of Rasch Analysis

Essentially the process of Rasch analysis involves comparing the structure of one's data against the Rasch model. The model assumes unidimensionality of the underlying construct being assessed and will be demonstrated in a number of ways by the data generated through the assessees engaging in the assessment. Some items may be more difficult to pass, or be more difficult to rate highly on, than others. More competent students will tend to score more highly than less competent students and less competent students are likely to fail the more difficult items. Finally, any variations from this pattern may be attributed to problems in the measurement instrument, rater inconsistency or students' knowledge gaps, and will require investigation (Bond & Fox, 2001; Linacre, 2002).

The following description of the mathematical process used by Rasch analysis to compare one's data to the model is primarily derived from Bond and Fox (2001). The mathematical model is based on the probabilistic relationship between any item's difficulty and any person's ability, for example, all persons have a higher probability of passing an item of low difficulty. The difference between the item difficulty and the person's level of competency will determine the probability that he/she will gain a particular rating. The primary discovery of Georg Rasch in mathematically modelling this relationship was that it could be used to predict the relationship between item difficulty and person ability for any person. This makes the process independent of the sample that is tested against the model and is the source of Item Characteristic Curves (ICC) which predict the scores or ratings (Y axis) a person of a particular ability (person location on the X axis) should receive (the line in Fig 6[6]) and the scores they are observed to received (dots on the line).

Figure 6. Example of Item Characteristic Curve



To produce the ICC the mathematical model estimates the ability measure of a person by converting the person's percentage of ratings on the assessment (sum of the ratings given on all the scales divided by the maximum sum possible) into an 'odds of success' figure. Thus a raw score of 55 (rating 5 on all items on the assessment tool) represents 70% correct as a percentage from the maximum possible raw score of 77 (rating of 7 on all items) and is

[6] These results are for illustration purposes and will be covered in more detail in the section detailing the statistical analysis of the field trial data.

converted into an odds ration of 70/30. This can then be converted into natural log of these odds and represents the person ability estimate or Rasch score.

Rasch analysis of rating scales uses an equation that includes the number of response choices the rater has and introduces thresholds which are difficulty estimates where a person has a 50/50 chance of being rated as being in one category over another e.g. the probability of being rated a 2 or a 3. This probability is expected to increase as person ability increases and is a combination of item difficulty plus threshold difficulty and the person's ability. These thresholds are estimated across all items and applied to all items. A further development of the equation can be used to enable the thresholds to be estimated separately for individual items and is termed partial credit analysis.

This modelling formalises what can be identified in the raw data about person ability and item difficulty. For example, in Table 13, it can be seen that students 1 to 98 have patterns of ratings that indicate that Item B is more difficult overall than items A, C, and D. Of the 3 students featured, Student 3 has ratings that indicate that she has a higher level of competency. Students 99 and 100 are clearly being rated in a manner that varies from that which would be predicted by the ratings of the remaining 98 students and this anomaly would be highlighted by fit statistics also calculated by the Rasch model.

Table 13. Example of the Calculation of Item Difficulty, Person Measures and Fit Statistics using Mock Data

	Items and ratings			
Student	A	B	C	D
Student 1	1	1	1	1
Student 2	2	1	2	2
Student 3	3	2	3	3
		(Students 4 to 99)		
Student 99	1	2	1	1
Student 100	2	1	3	6

Thus, the assumptions described and resulting mathematical model provide the basis for calculation of item difficulties, person ability levels, fit statistics, and category functioning (Bond & Fox, 2001). Each of these statistics has been used in the current study so will be described in more detail below:

1. Item difficulties are assigned through using the ratings provided by the CEs of students' performances and evaluating how hard it is to rate highly on this item.

2. Person ability, or the competence level of students, is estimated by looking at how highly the person is rated on the items. High ability students will rate higher on more difficult items and vice verse. This process converts the raw score summary of ratings on the test to a linear interval measure (as described above). The unit of measurement that is produced is termed a 'logit' and exists on an interval scale that quantifies both the ability of the person being measured and the difficulty of the items by which their ability is being measured.

3. Fit statistics, which identify how well the data fits the model, are calculated. The Rasch model is probabilistic in that it estimates the likelihood that a person of a given competence will achieve a certain rating on an item of known difficulty. In this way it is able to flag responses that don't 'fit' the predicted pattern for both items and persons. For example, fit statistics for items will be unacceptably high if highly competent students frequently receive a low rating on an easy item, suggesting that an item may be unclear or misunderstood by the rater or may contravene the assumption of unidimensionality. Fit statistics for persons will also indicate when ratings for a specific person do not fit the pattern expected for that person's ability level, allowing for a reasonable amount of variation. Other information is also provided by the analysis such as the standard error of the estimated person abilities and the item difficulties, which identify the degree of accuracy of the measurement (or how fuzzy or wide the lines of the ruler are). Two overall statistics for the total assessment are calculated. First, person reliability, which provides information as to how confident the test developer can be that a line of enquiry has been developed that identifies persons of different levels of competence. Second, item reliability, that indicates the degree of confidence the test developer can have as to whether the test contains items of different degrees of difficulty that provide a good description and hierarchy of competence (Linacre & Wright, 2003). Item functioning with different subgroups within the sample can be examined and is termed Differential Item Functioning (Andrich & Sheridan, 2004a).

4. Category statistics can also be calculated and comprise a number of informative statistics regarding how well the categories generated from the VAS measures function within the data. This includes fit statistics that indicate if there is too much or too little randomness in the category use. Information on how probable it is that a person of a particular competency level will fall into a particular category is also provided. For example, if a category is rarely observed in the data it may not represent

a meaningful ‘slice’ of the continuum of competence e.g. represent too narrow a section of the variable (Linacre, 2002).

6.3.3. Rasch Analysis of Rating Scales

Given that the focus of this research is on the validity of rating scale data, a brief explanation of Rasch analysis of rating scales follows and will clarify the framing of the research questions. Rasch analysis was originally applied to the analysis of dichotomous data such as pass/fail scoring on items comprising educational assessments and has been extended to cover polytomous data such as rating scales, responses that could be given partial credit (graded as to the quality of the answer), and testing situations where facets other than just the person or item need to be measured (Bond & Fox, 2001).

Rating systems are generally designed to produce ordered responses either by providing categories that represent increasing amounts of a particular characteristic or a visual cue such as the VAS in this research where the rater is asked to place a mark representing relative ‘amounts’ of competence. Rasch analysis is able to assist with two priority issues for this research. First, to analyse the functioning of the VAS scale used in the research as Linacre (2002) points out:

> “Since the analyst is always uncertain of the exact manner in which a particular rating scale will be used by a particular sample, investigation of the functioning of the rating scale is always merited.” (pp. 85)

As will be described in the section on research methodology, the raw data generated from the VAS scales is provided in the format of 100 equal sized units of measurement derived from both hard and electronic responses (see Section 6.4.5). Intuitively, it is extremely unlikely that raters are able to make 100 distinguishable and evenly separated distinct judgements about students’ competency. Indeed, a number of researchers have identified that measurements from a VAS do not represent equally spaced interval data (Cook et al., 2001; Linacre, 1998b; Munshi, 1990) and that raters are not capable of making more than 10 meaningfully different categories of judgement (Linacre, 1998b; Thomee et al., 1995).

Thus, an essential issue for this research is to identify exactly what the VAS ratings actually mean. This includes questions such as how many discriminations or categories are the CEs reliably making regarding competence? Are the critical points for moving from one level of discrimination to another along the VAS (thresholds indicating the move from one

category to the next) the same size and distance apart for each of these categories? If they are not the same size, what VAS scores are represented in each category?

Chapter Seven will describe the Rasch analysis process required to determine whether the data generated by the VAS measurements in this research can be treated essentially as 100 categories of judgement and, if not, how many categories should it be recoded into and where should the boundaries be drawn for each category. These questions are integral to the structural validity of the tool as the scoring structure of the assessment tool must match the way in which it is used in reality when rating the degree of competence of students' performances. This will ensure that data generated by the VAS scoring strategy is meaningful to the domain of competency and is given as much and no more credibility than it deserves and sheds light on how competency is perceived by CEs.

The second step, and the ultimate goal of this research, is to then derive meaningful measures of student competency on the basis of the ratings they receive from their CEs on the VAS.

6.3.4. Analysis Plan

The methodology of the field trial was expected to yield both qualitative and numerical data. The numerical data was subsequently converted into quantitative total score, or person measure, through the application of Rasch analysis. Qualitative data was required in the form demographic information regarding the sample with a view to enabling evaluation of the rating scale data from the assessment tool in relation to validity aspects such as generalisability and substantive validity. Qualitative data regarding the external validity of the assessment tool was also sought in the form of feedback from CEs and students who used the research tool during the field trial.

However, the evidence regarding the advantages of using Rasch analysis to evaluate the validity of the assessment tool via the numerical data yielded by the rating scales led to reframing the sub questions and hypotheses required to address the primary research question regarding the validity of the assessment process. The final list of questions for statistical assessment of the data, as it relates to the validity of the research assessment tool, is described in Table 14. The first two questions required answering sequentially as they determined the process by which the data would be analysed to answer remaining questions.

Table 14. Research Questions and Hypotheses and Related Validity Categories

Research Questions and Hypotheses	Messick's validity category
1. What is the fidelity of the scoring structure of data from the VAS to the domain of competency?	
Sub question 1: How many categories of competence did CEs discriminate on the VAS in an unambiguous and ordinal manner, as determined by Rasch analysis of the rating scale data and evaluation of the resulting categorical statistical indicators? Sub question 2: Are the critical points for moving from one level of discrimination to another along the VAS the same size and distance apart for each of these categories? Sub question 3: What VAS measurements are represented in each category?	Structural
2. Does the format of the assessment tool (electronic or hard copy) affect the way in which it is rated?	
Hypothesis 1: There will be no significant difference between ratings provided on the hard and electronic copies for each competency rated on the assessment, as determined by Differential Item Functioning (DIF) analysis for these two subgroups of data.	Generalisability
3. Does the measurement tool assess a unidimensional trait?	
Hypothesis 2: The assessment tool will measure a unidimensional trait as indicated by item fit statistics falling in the range of .8 to 1.2	Substantive
4. Does the assessment tool assess a range of ability levels?	
Hypothesis 3: The assessment tool will yield person measures ranging from at least – 3 to + 3 logits. Hypothesis 4: Person measures for more experienced students will be significantly higher than for less experienced students Hypothesis 5: Person measures will increase longitudinally for those students who have more than one assessment over the field trial period Hypothesis 6: Those students who are identified by their CEs as having performance levels on one or more competency that put the student 'at risk of failing their placement' will have person fit statistics that indicate variability in their performance (greater than 1.8) OR lower person measures compared to their similarly experienced peers.	Substantive
5. Can the person measures and their interpretations be generalised to different populations of students and workplace settings?	
Hypothesis 7: There will be a significant and strong positive relationship between students' person measures based on ratings provided by two different CEs working with the student in the same workplace at the same time. Hypothesis 8: There will be a significant and moderate positive relationship between students' person measures based on ratings provided by two different CEs in two different but concurrent	Generalisability

workplaces. Hypothesis 9: Item and person reliability measures (analogous to Cronbach's Alpha) will be greater than .80.	
6. How robust is the assessment tool?	
Sub question 4: Can one global overall rating of performance be used in place of specific competency ratings? Hypothesis 10: There will be no significant difference between the ratings given by experienced and inexperienced CEs to students with similar person measures as indicated by a) DIF analysis according to self rated degree of experience; and b) Analysis of Variance (ANOVA) between subgroups of clinical educator experience. Hypothesis 11: There will be no significant difference between students' person measures or performance on specific competencies according to the program they attend.	Generalisability
7. What is the level of satisfaction with the research assessment tool amongst students and CEs?	
Sub question 5: How satisfied are CEs and students with the research assessment tool, both overall and with regard to specific features of the assessment tool?	External

6.4. Research Method

The following section will describe the process undertaken to assess the validity of the assessment tool through collecting data from an extensive field trial of the tool with speech pathology students and their CEs across Australia. Piloting is usually recommended to identify and eliminate any problems with practical aspects of research procedures such as wording of research instructions or the assessment tool and resources themselves. However, a formal piloting phase was not included in the research method for three reasons. First, the development process was highly consultative and the research group were confident that this process was sufficient to ensure that assessment materials were comprehensible. In addition, materials produced to support the research process such as the research instructions and the online database were reviewed by several volunteer CEs as well as all members of the expert group to ensure their clarity.

Second, the potential pool from which data could be collected was already small and the research group was unwilling to sacrifice any data that could be useful for the validity analysis of the tool, the critical focus of the research. Third, it was anticipated that the advantages accrued through a piloting phase did not warrant the disadvantages created by the extended time frame this would impose upon the research process. This decision was justified

post hoc when the data collection process was subsequently extended to ensure that sufficient data for analysis was collected. The recruitment procedure, research materials, and methodology, including information on data collection, management, and tools used to analyse the data will now be detailed.

6.4.1. Recruitment Procedures

A three tiered approach to recruiting research participants was employed. First, the researcher approached each Australian speech pathology program with information regarding the research and solicited their support and, where required, assistance in recruiting their students. All eight programs expressed interest; seven consented to support the research through facilitating access to the students who would be having work placements throughout the field trial phase, with the eighth program facilitating access to one cohort of students towards the end of the field trial.

The second step involved presenting the research information to the students and asking them to indicate their consent via the return of completed consent forms. This consent enabled the students' placement details to be released to the researcher, indicated the students' agreement to being assessed with the trial tool, and for this assessment information to be provided to the researcher.

Information about the research was presented to students both verbally and on a written information sheet that had been approved by the relevant university ethics committees. The researcher approached students at Flinders University in person and students at other universities were either approached by members of the expert group (The Universities of Sydney, Newcastle and Charles Sturt) or by staff from their own university who had been briefed by the researcher (Macquarie and La Trobe Universities and The University of Queensland). All those involved in consenting students were provided with briefing notes that detailed the information to be presented and the consenting procedure to be used (Appendix 17). Once collected, student consent forms were posted to the researcher via a reply paid envelope.

The third step involved the researcher collating a list of those students who had returned signed and witnessed consent forms for each university. These lists were forwarded to the participating programs, which then provided contact information for the students' workplace placement(s) to the researcher. Finally, a research package was posted to each student's CE(s) who was invited to participate in the research through submitting an assessment of the student

via the research assessment tool. CEs consented to participate by either returning a signed consent form or by implication through returning a completed research assessment tool.

6.4.2. Research Materials

A package of research materials was provided to each CE who had one or more consenting students placed with them and consisted of information about the research and instructions on how to participate, assessment tool, resource materials, feedback questionnaire, and information on how to return assessment data and feedback questionnaires (Appendix 18). Participants were directed to contact the researcher for a copy of the CBOS as this is widely available at speech pathology workplaces, from the Speech Pathology Australia website, or could be forwarded in electronic or hard copy by the researcher. The other components of the research package are described in more detail in the following sections.

6.4.2.1. Demographic data

Demographic data was collected to both describe the sample of CEs and students and to provide information on variables that were thought to be relevant to student performance and CEs' use of the assessment tool. Thus information was collected from CEs regarding how many years since they had graduated and how many years they had practised as a speech pathologist. They were also asked for information to describe their experience as CEs, including the total number of students with whom they had worked and what universities and year levels these represented, and were asked to rate their level of experience as a CE. In addition, CEs were asked to indicate if they had received any training on assessment of students and to rate their familiarity with CBOS. Students were asked to provide information regarding their year level and university, hours of clinical experience, their familiarity with CBOS, and information on the placement type and client group they were working with when assessed.

6.4.2.2. Assessment Tool

CEs had the choice of returning the research data via a hard copy (paper) version of the assessment tool or via the online system proposed during Phase 1 of the research process. Detailed instructions and materials were provided to guide CEs as to how to carry out the assessment (Appendix 18).

The hard copy of the assessment tool is reproduced in full in Appendix 15 and its original format described. Mid placement assessment involved the CEs placing a vertical mark on the VAS at the point that represented the students' level of competency on each element of each competency (see Section 5.7.3 for an explanation of units and elements of competency). An overall rating for the whole unit of competency was made on a VAS on the page following the element ratings. End placement ratings required CEs to give the student a single overall rating for each unit of competency. This mark was the second mark made on the VAS, the first having been made as an overall rating on this unit at mid placement assessment. CEs were also required to indicate an overall level of competency for the students' total performance at both mid and end assessment. Space for comments was also provided.

Examples of the online system are provided in Appendices 19, 20, 21, 22 and 23. The online assessment format comprised a website that was accessed by typing in the URL into a standard web browser program available on any internet connected computer. CEs were then able to select from the web page whichever option they needed (Appendix 19). Documents available from the web page included demographic forms for students and CEs, all the resource materials available in the Assessment Resource Manual, and the assessment database.

CEs were provided with a user name and password that provided access to an assessment form for each student they were assessing (Appendix 20) as well as demographic form for themselves (Appendix 21). Students were given a user name and password that accessed their demographic form only (Appendix 22) so that they could not alter any ratings on their assessment forms. In addition, the researcher had password protected access to administration functions of the online assessment and could enter and link CEs with their student(s), access at the students' assessment forms at any time, and identify who had completed their mid and /or final assessments (Appendix 23). Students and CEs were represented by unique username and password combinations with no identifying information included online.

Once logged in to their own assessment page (Appendix 20, Step One), CEs were able to select from the list which student they wished to assess (Appendix 20, Step Two). This screen also let CEs know which mid and end placement assessments had been completed for which students. Once a particular student's user name was selected, that student's assessment screen would appear which listed the competencies to assess at the unit level (Appendix 20, Step Three). CEs would then select the first unit he/she wished to rate and this screen would appear (Appendix 20, Step Four). At mid placement assessment a rating scale appeared for each

element as well as an overall rating for the element, at end placement assessment the rating was for the whole competency only.

Once the mid placement assessment was completed, the CE was required to rate the students' performance overall (Appendix 20, Step Five). Once this was completed the end placement assessment competencies would then be listed on the student's assessment screen and the mid placement list moved to the bottom of the screen (Appendix 20, Step Six). This did not occur until CEs had indicated 'yes' to the question as to whether the assessment was complete and the system did not allow CEs to indicate 'yes' unless all competencies were either rated, or the 'not observed' option was selected and all other questions answered.

While CEs were assessing the student there was a side bar always present on the right of the screen that provided information on how to navigate through the system, and a hot link to resource material specifically linked to that competency e.g. the option of viewing the detailed application of the Behavioural Descriptors for that particular competency as a guide to the behaviours that might be observed for the 3 levels of skill. CEs simply rated the student by placing their mouse on the 'slider' of the scale and moving it to the position they felt represented the student's skill level. At End Placement assessments, once CEs selected a particular unit of competency to rate, they would find the mid placement mark they had made for the overall rating on each competency represented by a blue line above the rating scale on the end assessment rating page for each unit and for the overall rating (Appendix 20, Steps Seven and Eight). This provided a point of reference against which to assess any progress in the second half of the placement. Once the CE indicated that the final assessment was completed and returned to their index page, those students whose assessment was finalised would be indicated by a change in colour (Appendix 20, Step Nine).

6.4.2.3. Feedback Questionnaire

Feedback was requested from students and CEs regarding their experience of the assessment tool to assist in identifying whether the assessment validity was safeguarded through active engagement with the assessment tool (see Section 3.3.2.). In addition, information on the students' perceptions of the assessment was sought as this influences the quality of the learning that occurs, an aspect of assessment that should not be neglected (Maclellan, 2001). Questions dealt with users' experience of the research assessment tool, its perceived face validity, and factors that might affect its reliable and valid use. The content and

method used for the questionnaire as well as the results and interpretation are detailed in Chapter Eight.

6.4.3. Research Process

The research process primarily consisted of asking CEs to use either the hard copy or online research assessment tool to assess the student, prior to using the usual university tool at both mid (half way through the placement) and end placement assessment occasions. They were also requested to not alter ratings after having carried out the usual university assessment. Details and examples were provided as to the process of making ratings on the hard copy of the assessment tool.

Thus participation generally required CEs to fill out two assessment forms at each assessment event[7]. In addition to providing demographic information, CEs were also invited to fill out their section of a feedback questionnaire regarding their perceptions of the assessment tool. At the end of the hard copy tool, and also on the feedback questionnaire, they were asked to indicate how long it took to complete the mid and end assessment with the students.

A number of options for student involvement in the research were considered as their opinion of the assessment tool was valued by the research team. However the research team were aware that the process of data collection was dependent on CEs who would have varying philosophies regarding student involvement in assessment. In the opinion of the team, it was not possible to control the impact of this factor upon the degree of involvement students had in the assessment process without creating a separate study, which would decrease the amount of data available for the validity study. It was decided to recommend that CEs involve the student(s) in the assessment as per their usual practice and identify the level of involvement in the space provided at the end of the assessment tool. CEs were also asked to manage the student(s)' participation in the research process including ensuring that each student supplied demographic data (either in hard copy or online) and was invited to fill out the student section of the feedback questionnaire.

A number of students had more than one CE in the same placement or different placements. Usually CEs who work with the student in the same placement collaborate to provide a joint assessment of the student's performance. However all joint CEs were

[7] Some Clinical Educators at The University of Sydney were the exception to this during semester 1, as the research assessment tool was nominated as the assessment component of placements for particular year groups.

requested to carry out the assessment on the research tool without consulting with the other CE regarding their ratings on the assessment tool. This would enable this data to be used to assess the validity of the assessment tool with regard to the similarity of judgements or ratings between two CEs evaluating the same student.

6.4.4. Data Management and Collection

Data management and collection had several facets including collecting, storing, and managing data necessary for successful progress of the research data collection process itself. Research data collected for analysis included VAS measurements, CE and student demographic data, and questionnaire responses.

The recruitment process and data collection was managed in a Microsoft Access 2000 database designed specifically for the purpose. Data was stored on interlinked tables recording relevant information regarding placement sites, the CE(s) at these sites, student information including general identifying information, university affiliation placement dates, and CEs over the research period. Each student and CE were allocated an ID comprising of 2 letters with or without a number, and a 6 letter nonsense password generated by a random password generator (WinGuides, 2003).

The database program enabled the automated production of instructions for accessing the online version of the assessment tool that included the correct ID codes and passwords for the CE and each student thus avoiding errors likely to occur when manually collating this information. Information on student and CE pairing and placement timing and length to manage mail outs could be retrieved by formulating the appropriate queries in the database. The database itself was password protected and run on a personal computer with appropriate security software preventing unauthorised access to the database when the host computer was connected to the internet.

The online assessment tool included an administrator component accessible only by entering the correct name and password. The administrator function enabled the researcher to enter the unique ID and password for all students and CEs and to identify linked pairs. A computer program was utilised to semi-automate this entry to minimise errors. Thus, once CEs accessed the website and entered their ID and password, they could generate a specific online form on which to rate each particular student placed with them. Students were only able to access their demographic record and could not use this ID and password to access and modify the assessment data.

The online database was hosted by Portal Australia and protected from unlawful access by use of randomly generated passwords and state of the art virus protection or 'firewalls'. No identifying information was included in the online database, ensuring that the data collected online was meaningless to anyone except the researcher. The risk to the data through mischievous or intentional interference was assessed as being extremely low (Brett Kokegei, Portal Australia Inc., personal communication, December 4, 2002). Once the research was completed the database was converted to allow the researcher to enter hard copy measurements as numbers for each VAS as well as demographic data submitted by hard copy. VAS and demographic data were downloaded as a total group by Portal into a Microsoft Access 2000 database. This was done on several occasions so that missing CE and student demographic data to be identified and pursued. Student and CE responses to the feedback questionnaire, both ratings and comments, were first entered into a Microsoft Excel 2000 spreadsheet and subsequently into Microsoft Access 2000 database.

6.4.5. Measurement of Ratings

Ratings entered online were automatically converted into 100 unit measurements based on where CEs placed the sliding bar on the rating scale. However, hard copies required hand measurement to determine the point at which CEs had placed the mid assessment mark for 65 rating scales and the end placement mark for 12 rating scales, a total of 77 scales.

A set of 10 hard copy assessment booklets, representing 770 scales, were measured using electronic callipers that provided a measurement in millimetres to two decimal points. Each set was measured by the researcher and by a research assistant employed with grant funds. While measurements in millimetres to two decimal points provided a greater degree of accuracy than required, the callipers were easier to use and provided clearer measurement information than a ruler. Each person measured the rating scale by lining up the vertical mark on the left hand end of the VAS so that it created a straight line with the left hand edge of the calliper. The calliper was slid open until the right hand edge lined up with the middle of the mark made by the CEs and the measurement noted from the electronic display. If the mark was not exactly on the line, the measurement was taken from the point that was directly below the end of the mark if it was a single mark or where the two points crossed over if it was marked with an X.

Measurements were entered into a Microsoft Excel 2000 spreadsheet. Each measurement was converted into a 100 unit measurement as a proportion of the length of the original VAS.

This length was originally intended to be 100 mms but it was discovered that the printing process had lengthened the scale to 101 mms and that some photocopied versions of the scales were very variable, measuring up to 105 mms. The conversion to 100 units was made to 5 decimal points, and the two measurement sets compared to each other by subtracting one from the other and then rounding the remaining number up or down to identify any differences in measurement equal to or greater than 100th of the scale.

In total, 8 scale measurements out of 770 scales jointly measured were found to differ by greater than 100th of the original VAS length, representing 99% agreement between the two measurers. Examination of the mismatches in ratings indicated that the research assistant was more accurate in her measurements (1 error compared to the researcher's 4 errors), an accuracy rate of 99.9%, and that some marks required consensus. Thus the research assistant was employed to measure the remaining hard copies and consulted with the researcher to resolve unclear markings.

6.4.6. Data Analysis

Data to be analysed included demographic information such as CE and student characteristics, as well as the ratings generated by the VAS. Descriptive statistics were entered and analysed using SPSS Version 12 (SPSS, 2003) as well as parametric statistical analyses on interval data (measurement of student competence) generated through Rasch analysis. Rasch analysis was conducted using two programs. Bigsteps (Linacre & Wright, 1998) was used for the majority of the analysis and RUMM (Andrich & Sheridan, 2003) used to carry out Differential Item Functioning analyses, a feature unavailable on Bigsteps.

6.4.7. Summary

A research methodology was designed to collect demographic, process, feedback, and rating data on the use of the research assessment tool via hard copy and online formats. The method endeavoured to collect sufficient data to enable a thorough evaluation of the tool's validity and to safeguard students' rights to consent without coercion and to have their assessment data treated confidentially.

6.5. Description of Sample

As described in the methodology, a convenience sample was used and the following section provides a description of the data received and the participants in the research. Note however, the need for this sample to exactly mirror the total population of students from which it is drawn is not as critical when using Rasch analysis approaches as compared to CTT because the measures generated by Rasch analysis are not sample dependent.

6.5.1. Consent Rates

Seven of the eight existing university programs agreed to assist in approaching all their students at the start of the field trial with Curtin University approaching students late in the trial phase but without success. Student consent rates varied significantly among programs and are summarised in Table 15. Only one student withdrew consent part way through the research.

Table 15. Students Consenting to Participate in Field Trial

University	2nd yrs	3rd yrs	4th yrs	2nd yr Post Graduate	Total consent
University of Newcastle	19/30	19/32	3/25	N/A	31/60 (52%)
Flinders University	N/A	19/21	22/22	N/A	41/43 (95%)
Macquarie University	N/A	N/A	N/A	16/17 (94%)	16/17 (94%)
Charles Sturt University	N/A	14/34	13/32	N/A	27/66 (41%)
The University of Sydney	36/55	62/74	41/ 60	N/A	139/189 (74%)
University of Queensland	Information not supplied				17
La Trobe University	Information not supplied				11
TOTAL	Consent rate averaged 67.7% for universities excepting The University of Queensland and La Trobe University				282

6.5.2. Returns

By the end of the trial period approximately 563 CE/student pairs had been invited to participate in the research. Return rates are approximate as the researcher was advised by some CEs that the student was no longer placed with them. Overall 321 assessment formats were returned to the researcher, a return rate of at least 57%, 236 were received electronically and 85 were received as hard copies. Each assessment event (321) represents a unique student and CE combination for which data was been collected.

This data represents assessments from 219 different students by 107 different CEs. This constituted an overall consent rate for CEs of 44% given that 246 were approached. An estimated[8] 58 CEs employed or part funded by university programs were invited to participate, with 47 returning 238 assessments (74% of the data). Sixty field educators returned 83 assessment formats, representing 26% of the data. The majority of CEs associated with universities submitted their assessment data electronically (44 or 94%) with relatively fewer field CEs (26 or 43%) submitting electronically.

6.5.3. Nature of Data Returned

End placement data was included in 301 of the 321 assessment events returned. The convenience sampling strategy resulted in a number of students having assessment data submitted by 2 CEs who were working with the student concurrently as well as 2 or 3 sets of data being submitted for some students over the 10 month trial period. Twenty students had 2 assessments each submitted by CEs providing a placement in the same or very similar workplaces simultaneously. Forty-four students have 2 assessments each from CEs working with them at the same time but in different workplaces. Twenty-nine students had data collected longitudinally.

6.5.4. Clinical Educator Characteristics

Demographic data from the CEs was not received from all participants and a few had missing data. Of the 93 that provided information experience as speech pathologists varied from 1 to 33 years, with the average being 9 years, however, the amount of data submitted by

[8] The numbers are approximate as they are inferred from the mailing addresses for CEs as well as information from The University of Sydney and Flinders University regarding joint funded positions, and so may slightly underestimate the number of university funded CEs involved in the data collection.

speech pathologists with differing years if experience was variable (Fig. 7). The majority of data was submitted by CEs who had assessed more than 20 students (Fig. 8) and rated themselves at 5 on a 7 point scale of ‘experience as a CE’ (1 indicating their first student and 7 indicating they were very experienced) (Fig. 9). Data was submitted primarily by CEs who considered themselves to be moderately familiar with CBOS (5 on a 7 point scale) or very familiar (rating of 7) as illustrated by Fig. 10. Seventy nine (74%) CEs indicated that they had attended a workshop on clinical education which resulted in a similar proportion of the data (231 assessments or 72%) being submitted by CEs who had attended training.

Self rated experience correlated moderately with the number of students that CEs had assessed (Spearman’s rho = .55, p >.001). Given the different nature of placements in length and intensity, it is understandable that self rated experience may not have a 1 to 1 correlation with the number of students that CEs had supervised. CEs who have provided 9 x 10 week placements for 4 days a week in a field setting over 5 years may consider themselves highly experienced as a CE. On the other hand CEs employed by a university program for the first time to work with 15 students for a few appointments a week each may not rate themselves as being particularly experienced. Thus self rated experience could be considered to be a better indicator of experience than the raw number of students CEs have worked with in the past.

Figure 7. Years of experience of clinical educators participating in field trial

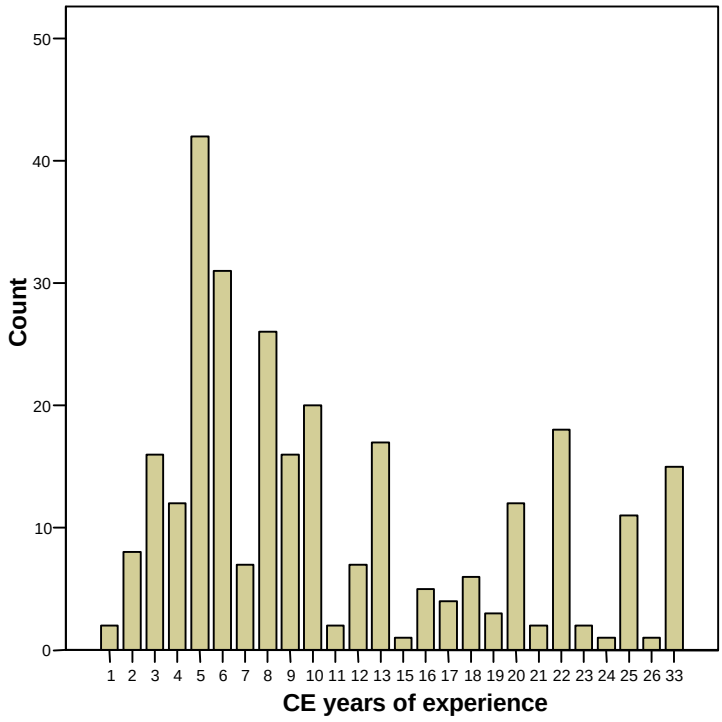


Figure 8. Total number of students supervised over the career of clinical educators participating in field trial

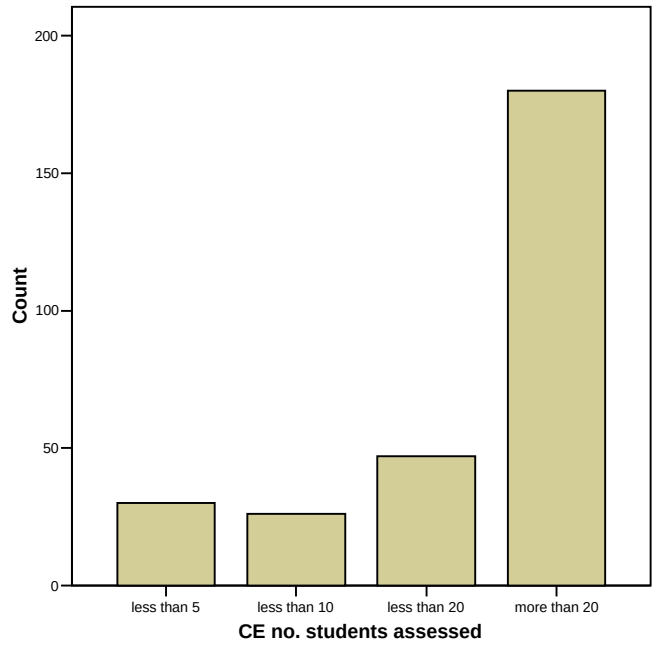


Figure 9. Self rated experience level of clinical educators participating in field trial

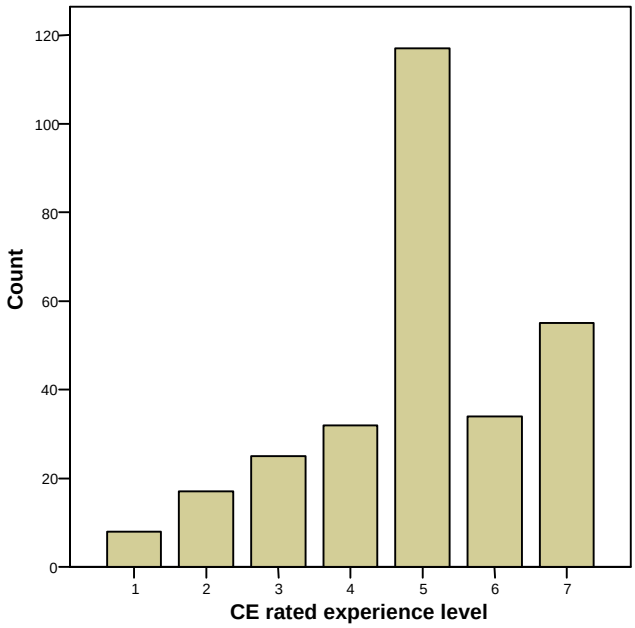
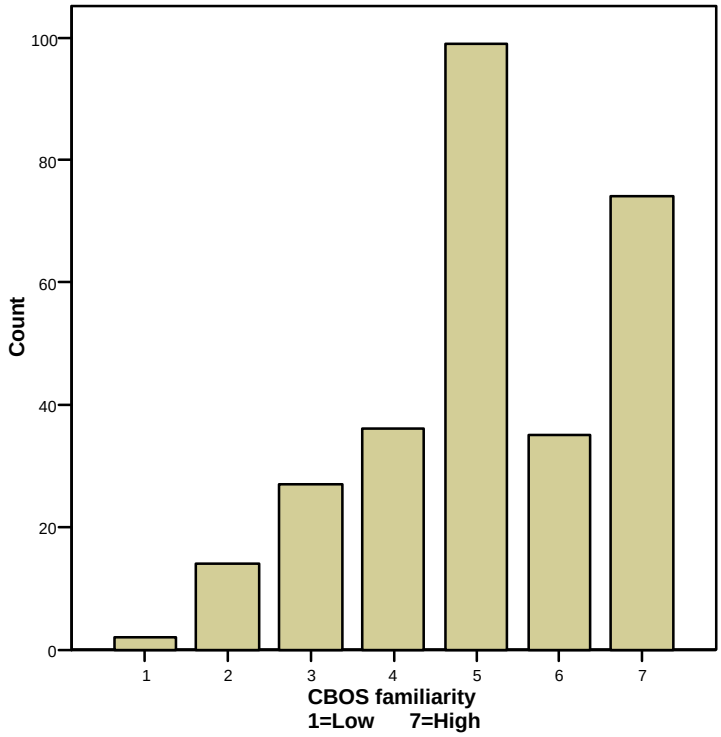


Figure 10. Self-rated familiarity with CBOS by clinical educators participating in field trial



6.5.5. Student Characteristics

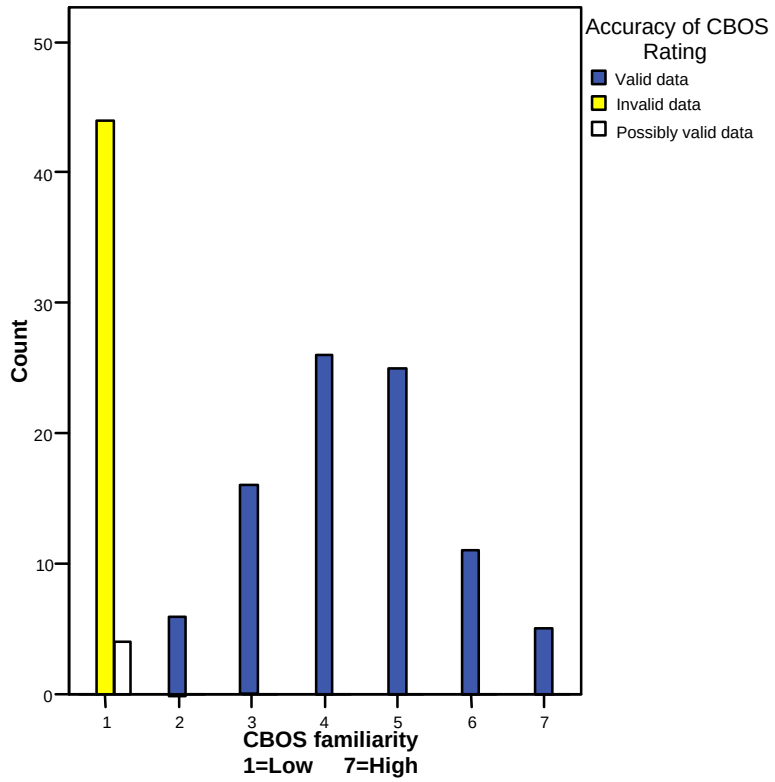
As would be expected by the consent rates described above, the majority of students for whom data was collected were from The University of Sydney (64.8%) however 35.2% of the data was collected from students from other universities (Table 16). Demographic data was returned by 137 students but not all data was complete thus the demographic information is not descriptive of all students who had data returned for them by CEs. Demographic information was collected on students' familiarity with CBOS, type of placements they were in, client groups they served, the number of hours they had accumulated by the end of the placement, and how many days/weeks they were on work placement.

Table 16. Universities Represented in Field Trial Data

University	N	Percent
Charles Sturt University	12	4.0
Flinders University	42	14.0
La Trobe University	5	1.6
Macquarie University	14	4.6
The University of Newcastle	24	8.0
The University of Queensland	9	3.0
Subtotal	*106*	*35.2*
The University of Sydney	195	64.8
Total	301	100.0

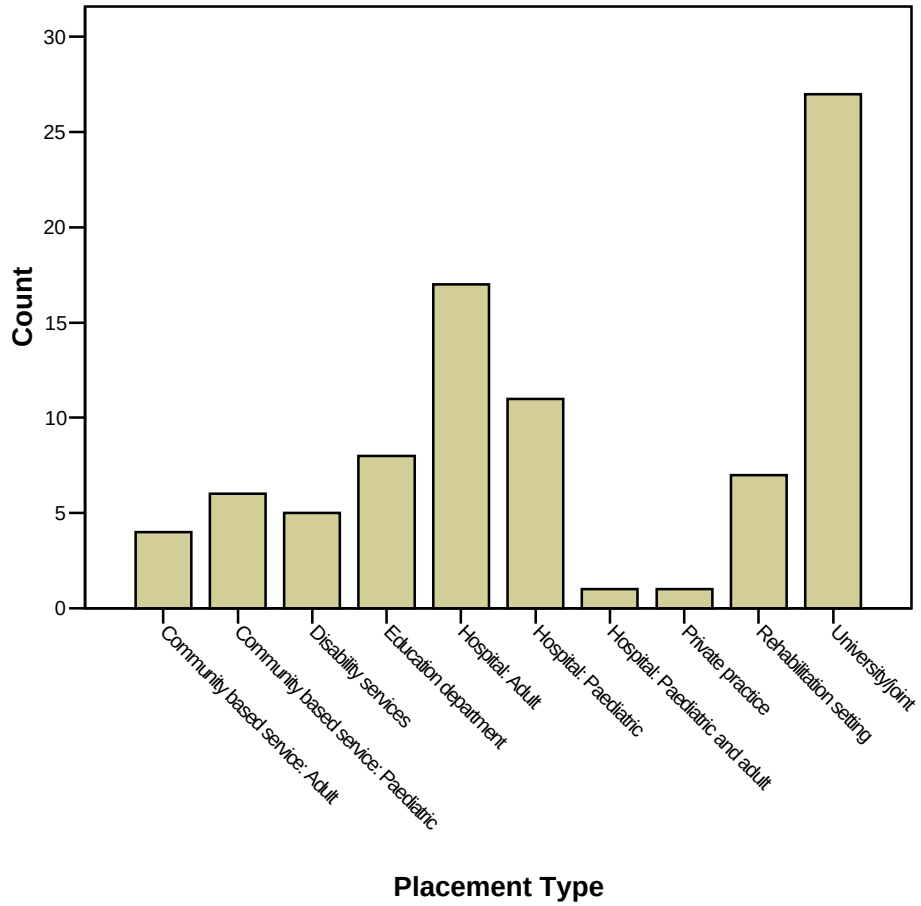
A number of difficulties were encountered when collating this data. First the database that data was entered into was configured so that fields such as CBOS familiarity returned a default selection of '1' if the students did not select/provide a rating. Examination of the data suggested that it would be appropriate to disregard an entry of '1' if no placement information was provided in the question prior to this as it appeared that students sometimes only entered their university information and the amount of experience they had in the placement (hours, days and/or weeks). A few students did provide other information on and the '1' entered in the database for their CBOS familiarity could represent an actual self rating of 1 (rather than an unselected default setting of 1). Fig. 11 summarises this information.

Figure 11. Self rating of familiarity with CBOS by students participating in field trial



Information on the placement was provided by some students but unfortunately the database configured by Portal omitted information on the client group. However, given so few students completed this information and concerns over verifying its accuracy, it was not considered valid to use either of these variables to define data for analysis. However, where reported by students, quite a variety of placements were represented, offering some reassurance that the validity of the assessment tool was being tested against a range of placement types (Fig. 12).

Figure 12. Placement types reported by students participating in field trial



However, a critical variable for analysis was the amount of direct client contact each student had experienced by the end of the placement. Some universities require students to document how much face-to-face time they spend in service provision, either directly in client contact or in client related activities such as case meetings, administration time is not counted. However, even for those universities who require their students to keep a total of direct client contact hours, a substantial amount of this data not provided. In addition, scanning the data indicated that many students misunderstood the question and did not provide an accumulated number of hours. Thus it was necessary to ensure either verified hours data was entered or to estimate it where it was not available. The University of Sydney and The University of Newcastle were able to provide the number of hours students had recorded for each placement, signed off by each CE and registered with their program. This accounted for 232 sets of data (hours could not be provided for 3 students from these universities).

For the remaining students, with the exception of La Trobe University (representing 5 sets of data), universities provided information on the structure of placements and how many each student had completed which provided a basis for estimating the number of hours students

had spent in direct client contact at the end of each placement. Some hours were able to be identified from the nature of the placement e.g. 1x 1 hour client session per week for 13 weeks. Where the hours of experience could not be predicted as easily they were estimated as being 2 hours per day for a full day 2nd or 3rd year undergraduate/1st year postgraduate placement and 4 hours per day for a full day 4th year undergraduate/2nd year post graduate placement.

This ratio was arrived at through the following process. Using the combined research group, the number hours students were likely to accumulate in direct service delivery to clients in an average per day of clinical experience was estimated. The average number of hours accumulated per placement day and per year level of student for those students who had this information available was then examined. Eleven students from The University of Newcastle accumulated an average of 2 hours per day for their 2nd year placements and the two 4th year students averaged 3.78 hours per day. Twenty-seven 4th yr students from The University of Sydney students averaged 3 hours per day for their placements, ranging from .9 to 6 hours per day. However, other university clinical coordinators agreed (Flinders, Macquarie and Charles Sturt Universities) with the estimated total of 4 hours of direct client contact per day for their final year students.

There were some total hours reported by students that could be compared to the estimated hours. One Flinders University student reported that 320 hours had been accumulated by the end of her final placement compared to an estimated 336 hours (see Table 17 for estimation totals). A second Flinders University student reported 296 hours as compared to an estimated total of 336. A Charles Sturt University student reported 250 hours as compared to an estimated 288 hours and a second Charles Sturt University student reported 200 as compared to an estimated total of 208.

It was clear that direct client contact hours accumulated per day of placement varied widely according to the ability levels of the students and the complexity of the placement and probably the teaching style of the CE. However, overall the estimation of 2 hours per day for ‘junior’ students and 4 hours per day for ‘senior’ students seemed a reasonable ‘rule of thumb’. Table 17 outlines the estimated hours entered in the data for students for each placement combination, year level, and students.

Table 17. Calculation of Estimated Hours of Student Experience to Classify Field Trial Data

University Program	Junior Student 1 day = 2 hours (est.)	Senior Student 1 day = 4 hours (est.)	Est. Total Hours
Charles Sturt University	24 days x 2 hrs = 48 hrs (Year 2 & 3)	80 days x 4 hrs = 320 hrs (Year 4)	368
Flinders University	24 days x 2 hrs = 48 hrs (Year 3)	72 days x 4 hrs = 288 hrs (year 4)	336
Macquarie University	28 days x 2 hrs = 56 hrs (Year 1)	40 days x 4 hrs = 160 hrs (Year 2)	216
The University of Queensland	2 hrs appts x 11 days = 22hrs 3 hrs appts x 12 days = 36 hrs 3 hrs appts x 12 days = 36 hrs 24 x 2 hrs days = 48 (Year 2 & 3)	24 days x 4 hrs = 96 hrs 3hrs appts x 12 days = 36 hrs (Year 4)	274
	Other calculations of hours		
La Trobe University	Estimation information and documented hours not available		
The University of Newcastle	Documented hours available		
The University of Sydney	Documented hours available		

Note. Estimated hours do not include observation or placements in 1st yr undergraduate programs.

Once student hours were identified or estimated, these hours totals needed to be usefully grouped to enable comparison with some aspects of the data. The configurations of the courses and related hours per placements were examined, and it was determined that grouping students by 3 levels of experience was likely to yield the most useful comparisons. These were identified as being 'beginner' (0 to 80 hrs), 'intermediate' (81 to 180 hrs), and 'advanced' (181+ hrs). Table 18 illustrates that these groupings tend to capture students at similar points of progression through their courses. However, this is an inexact estimate only as some students may have accumulated more hours than predicted by this table due to repeated or extended placements due to concerns regarding their development of competency.

Table 18. Student Progression Through University Programs in Relation to Groupings of Estimated or Actual Hours

University	0 to 80	81 to 180	181 to 300+

Charles Sturt University	2nd and 3rd year placemts.	4th yr 1st placemt	4th yr 2nd and 3rd placemt
Flinders University	3rd year placemts.	4th yr 1st placement	4th yr 2nd placemt
Macquarie University	1st year placemts	2nd year 1s and 2nd placemt	2nd yr last placemt
The University of Queensland	2nd year 1st and 2nd placemt	3rd year 1st and 2nd placemt	4th yrs 1st and 2nd placemt
The University of Sydney	2nd yr and 3rd yr 1st placemt	3rd year 2nd placemt., some 4th yr 1st placemt	Most 4th year 1st placemt., 2nd and 3rd placemt
The University of Newcastle	2nd year students, 1st 3rd year placemt	3rd year placemts	4th years placemts
La Trobe University	Information not available		

6.6. Summary

This chapter outlined the processes undertaken to determine the analysis required to assess the validity of the assessment tool (rating format and resource manual), the method by which data was collected to assess this validity, and the nature of the sample of CEs and students who participated in data collection. The next chapter describes the analysis undertaken of the numerical data, in the form of VAS measurements, to evaluate the assessment tool's effectiveness in measuring students' competency to practice speech pathology. Chapter Eight provides information on the method and results for the user evaluation of the assessment tool.

CHAPTER SEVEN

7. ANALYSIS OF ASSESSMENT TOOL DATA

7.1. Introduction

Rasch analysis is a relatively new strategy for evaluating the measurement properties of an assessment instrument and deriving internal measures from data. It has been used more commonly within educational and some health research fields but has not been previously applied to the assessment of competency in speech pathology (see Section 6.3.2. for a description of Rasch Analysis). Indeed, with the exception of a small study by Rheault & Coulson (1991) on using Rasch analysis in the design of an assessment tool of physical therapy students, Rasch analysis does not appear to be applied to the measurement of competency in workplace performance at all. Given the relative uniqueness and the reiterative, sequential analysis process required by this research, this chapter not only describes the process involved in analysing the data and results of the analysis, but also interpretations relevant to each subsequent stage of the analysis process. Chapter Eight will then describe the user evaluation of the research tool and process and then Chapter Nine will assess all the validity evidence for the assessment tool derived from the development phase and analysis of the field trial data.

7.2. Analysis of VAS Functioning

7.2.1. General Description of Analysis Process

The first task of analysis is to investigate whether the rating scale observations conform reasonably closely to the model proposed. Thus it must be determined that the ratings identify discernable degrees of a unidimensional trait that persons possess, of an infinite variable, in an unambiguous, ordinal fashion (Linacre, 2002). This process involves examining the statistics provided by the Rasch analysis and using them to identify if the categories function well and what action, if any, needs to be taken to optimise the rating scale's accuracy. Andrich and Wright (1994) agreed that the analyst should identify the number of categories that actually work in the data and then collapse the data into those groupings. This approach has been used frequently over the past decade to evaluate the functioning of a number of

rating scales e.g. Smith, Wakely, De Kruif, & Swartz (2003); Wright & Linacre (1992) and Zhu (1996); and Linacre has now published an article providing guidelines as to the appropriate decision making process to do so (Linacre, 2002).

The process of evaluating the rating scale, as an effective strategy for gathering data that can be converted into a measure in which we can have confidence, involves an iterative procedure of examining statistical indicators of rating category appropriateness and modifying the categories until they function in an unambiguous and ordinal fashion. As identified previously, the data from a VAS is essentially categorical, the 'categories' are generated by measurement of the VAS into equal sized intervals (usually 10 or 100) and are then examined and collapsed according to this procedure (J. Linacre, personal communication, January 2003). Once the analyst ensures that all items are aligned on the variable in the same direction (i.e. there are none with reversed polarity) detailed guidelines are provided by Linacre (2002) for optimising the rating scale and are summarised in Table 19 and explained in the following paragraphs.

First, a minimum of 10 observations is necessary for each category with 25 to 50 observations being required depending on the degree of stability of measurement required. The observations ideally should be distributed uniformly across the categories although unimodal or bimodal (peaking at extreme categories) can also be meaningful. The most problematic distribution is when there is a long 'tail' of categories that are relatively rarely used. Average measures should advance monotonically with categories. This means that the statistics generated for each category that compare the average measure of person ability for people receiving ratings in that particular category with the predicted average measure should be similar to each other and advance up the underlying variable.

Outfit mean squares of less than 2.0 are required to ensure that the data is not too variable as excessive randomness or 'noise' affects the measurement system the most severely. As described in Section 6.3.2., the Rasch model assumes there will be some degree of randomness or error within the data and specifies that it must be reasonably uniform indicated by mean square fit (infit or outfit) statistics of 1.0. A fit statistic of 2.0 suggests that there is a 100% more variation in the observed data than the Rasch model predicted (Bond & Fox, 2001). Thus a categorisation that yields outfits of above 2.0 is unacceptable and even stricter levels can be applied (Linacre, 2002).

Step calibrations must advance, these are the logit scores representing the transition points between categories where there is an equal chance that the person's ability falls into one

category or another, and so define the start and end of each category across the scale of person measures. Thus step calibrations are the point on the logit scale where the probability of a category being observed for a person of a particular ability level increases and becomes more likely (Linacre, 1999a). Increasing step calibrations indicate that as person ability scores progress up the scale (or the ICC such as the one in Section 6.3.2), each category in turn should become the most likely to be observed as indicated by higher and higher logit scores. If the step calibrations are disordered it indicates that there is a low probability of observing certain categories because of the way in which they are being used in the rating process.

Coherence statistics are also provided by the analysis and report the percentage of ratings that are expected to be observed in a category (as indicated by the person measures) compared to those that are actually in that category. At least 40% or above is acceptable if the data set is satisfactory in all other respects and ensures that ratings imply measures and measures imply ratings. The final requirement to make certain that an appropriate categorisation is operating is to identify that step difficulties advance by at least 1.4 logits but ideally not more than 5.0 logits. If the step difficulties increase by less than 1.4 logits it cannot be assumed that the rating scale is a series of clearly dichotomous items. Greater than 5.0 logits indicates that the category may represent a very wide range of performance with a ‘dead zone’ developing in the centre of this category thus losing precision. The pertinence of each of these guidelines is summarised in Table 19.

Table 19. Summary of Pertinence of Guidelines for Rasch Analysis of Rating Scale Categorisation

	Guideline	Measure stability	Measure Accuracy (Fit)	Description of this sample	Inference for next sample
Pre.	Scale oriented with latent variable	Essential	Essential	Essential	Essential
1	At least 10 observations of each category	Essential	Helpful		Helpful
2	Regular observation distribution	Helpful			Helpful
3	Average measures advance monotonically with category	Helpful	Essential	Essential	Essential
4	Outfit mean-squares less than 2.0	Helpful	Essential	Helpful	Helpful
5	Step calibrations advance				Helpful
6	Ratings imply measures, and measures imply ratings		Helpful		Helpful
7	Step difficulties advance by at least 1.4 logits				Helpful
8	Step difficulties advance by less than 5.0 logits	Helpful			

Note. Reproduced with permission from Linacre, 2002, pp. 104.

Other statistical information that should also be considered when carrying out an exploratory Rasch analysis includes how changes in the classification of the data or removal of suspicious data affects the item and person reliability statistics (D. D. Curtis, personal communication, August 2004). These statistics are analogous to Cronbach Alphas and can be interpreted in the following way (Bond & Fox, 2001). Person Reliability indicates the replicability of person ordering we could expect if this sample of persons were given another set of items measuring the same construct. It requires ability estimates to be well targeted by the items and a large spread of ability across the sample so a hierarchy of ability or development (person separation) on the construct exists. High person reliability means that we have developed a line of enquiry on which some persons score higher and some lower and that we can have confidence on the consistency of these inferences. Item Reliability describes replicability of item placements along the pathway if they were given to another sample with comparable ability levels i.e. the estimates of item difficulty would remain the same. High item reliability means that we have developed a line of enquiry in which some items are more difficult and some easier and that we can have confidence in the consistency of this inference.

Rasch analysis enables insight to be gained as to how the data cooperate to construct measures and the guidelines aim to assist in verifying and improving the functioning of the rating scale categories in the data collected (Linacre, 2002). The ultimate goal is to facilitate accurate measuring of the underlying trait of competence as:

> "Unless the rating scales which form the basis of data collection are functioning effectively, any conclusions based on those data will be insecure." (pp.104, Linacre, 2002)

7.2.2. Analysis of Rating Scale Functioning

7.2.2.1. Process: General Description

The data was analysed with Bigsteps 2.82, a DOS based Rasch analysis program (Linacre & Wright, 1998). The ratings (0 to 100) for each person for each Generic and CBOS competency were entered and statistics generated to examine the way in which categories function within the data as a whole. Analysis involved grouping and regrouping the ratings into categories based on the information generated the Rasch analysis until the final solution with the best functioning number and sizes of categories was identified. The range for acceptable fit statistics was selected to be .8 to 1.2, rather than around .5 to 1.5 or 2.0 as suggested by Linacre's guidelines, as the assessment represented a 'high stakes' event and more stringent expectations of the data were imposed as recommended by Bond & Fox (2001).

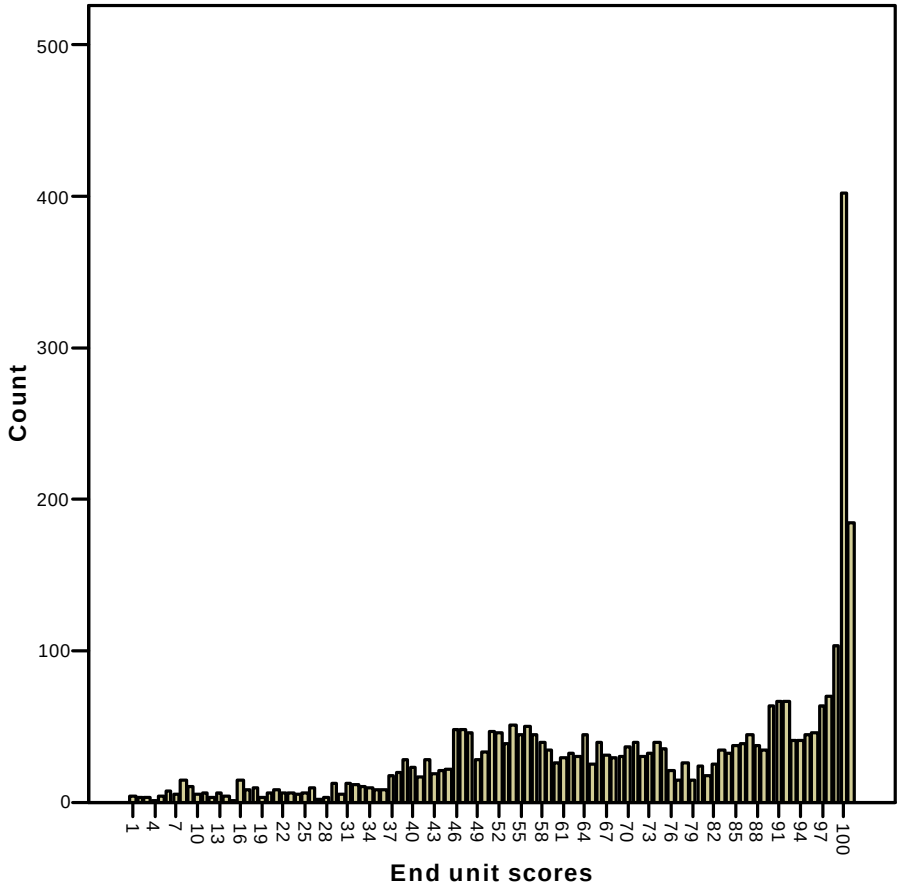
As described previously, the data comprised 100 units generated from the online and hard copy measurements and a final category of 'above entry level'. The final category representing 'entry level competence or above' was nominated as being represented by a combination of ratings of 100 and 'above entry level' and this category remained unchanged throughout the analysis. The remaining ratings of 1 to 99 were then analysed and treated as separate categories on the basis of the statistics provided by the analysis.

Ratings from 301 persons were entered in the data. Between 24 and 31 persons received the maximum or minimum score possible depending on the categorisation of the VAS scale used. These persons received ratings falling into the minimum or maximum category for all 11 items and thus are excluded from the calculations of summary and category statistics for person measures. Rasch analysis does this automatically on the basis that these persons possess an unquantifiable amount of the variable being assessed so that the rating lacks

sufficient measurement precision. This left ratings from 270 to 277 people available for person calibration calculations.

The ratings on the end placement assessment VAS were distributed in a bimodal pattern with a large ‘spike’ representing ratings of 100 followed by 101 (representing above entry level) and some clustering of ratings around the middle of the VAS (Fig. 13).

Figure 13. Distribution of students’ ratings across the 101 categorisation of the Visual Analogue Scale



7.2.2.2. Process and Results: Analysis steps

The analysis process to determine the number of categories functioning in the VAS was a reiterative procedure whereby the decision regarding where the VAS should be segmented was based on statistical information yielded by each preceding step in the analysis. This procedure is described below in a stepwise fashion, each step describing the process and result it provided that informed the next step in the process.

Step one

To ensure that each category had at least 10 observations the data was collapsed evenly into 33 groups derived from the first 99 ratings, plus the entry level group. This 34 categorization yielded a Person Reliability statistic of .97 and an Item Reliability statistic of .95 (Tables 1 and 2, Appendix 24) and other summary statistics such as item fit statistics fell within a reasonable range of .69 and 1.46, with a mean close to 1. However, there were insufficient observations for the rating category number 5 and the requirements regarding step calibrations, average measures, coherence, and fit statistics (Table 3, Appendix 24) were not met.

Step two

As the functioning of the categories were so disordered and there was no clear direction indicated by the statistics to guide the recategorisation decision, ratings of 1 to 99 were collapsed evenly to trial 16 categories and ratings 100 and above comprised the 17th category. The summary statistics for the whole sample when categorised in this way provided an improved item and person reliability statistic of .98 with item fit statistics ranging from .71 to 1.39.

The summary statistics (Table 20) for the 17 category solution suggested that further collapsing was required due to most step calibrations being less than 1.4 although this may have been permissible given the number of categories. However, one step calibration was disordered and the outfit statistics for categories 1 to 3 were too high. In addition coherence percentages were nearly all below 50% suggesting poor inferential power for this categorization.

Table 20. Rasch Analysis of Rating Scale Step 2: Summary statistics for 17 category solution

CATEGORY LABEL	OBSERVED COUNT	AVERAGE MEASURE OBS	AVERAGE MEASURE EXP	COHERENCE EXP%	COHERENCE OBS%	INFIT MNSQ	OUTFIT MNSQ	STEP CALIBRATN
1	23	-4.59	-5.26	64%	47%	8.09	3.78	NONE
2	40	-4.30	-3.98	74%	50%	1.12	1.35	-5.16
3	42	-2.90	-3.01	50%	50%	1.24	1.38	-3.46
4	42	-2.49	-2.46	45%	42%	.87	.93	-2.70
5	44	-1.98	-2.01	32%	29%	.88	.88	-2.27
6	59	-1.66	-1.59	20%	18%	.84	.80	-2.08
7	130	-1.25	-1.20	38%	34%	1.00	1.01	-2.18*
8	248	-.81	-.80	42%	44%	.84	.87	-1.65
9	274	-.38	-.37	38%	44%	1.08	1.11	-.69
10	210	.09	.11	36%	42%	.82	.85	.13
11	202	.66	.67	42%	39%	.76	.75	.43
12	205	1.41	1.34	46%	44%	.98	.99	.98
13	158	2.13	2.22	44%	37%	1.08	1.04	2.02
14	208	3.23	3.28	54%	42%	.75	.71	2.48
15	302	4.24	4.25	48%	60%	.72	.63	3.40
16	368	5.48	5.37	51%	58%	.86	.90	4.57
17	335	6.82	6.88	77%	58%	1.40	1.14	6.17

*step calibration out of sequence

Step three

Again, no clear guidance was available to indicate in which direction the categories should be collapsed so a 9 category solution was trialled. The measurements from the VAS that corresponded with these categories were as follows:

Category 1: 0 to 12.

Category 2: 13 to 25.

Category 3: 26 to 37.

Category 4: 38 to 50.

Category 5: 51 to 62.

Category 6: 63 to 74.

Category 7: 75 to 87.

Category 8: 88 to 99.

Category 9: 100, 101.

This recategorisation did not change the summary statistics for the whole data with regard to item and person reliability or item fits for the data but item fit statistics were improved, falling in the range of .76 to 1.22. The category statistics (Table 21) were much improved but the step calibration between category 3 and 4 was measured as only advancing by .44 logits contravening the requirement of advancing by around 1.4 logits. In addition both the infit and outfit statistics for category 1 were too high and the expected and observed coherence was poor. Coherence was also poor for category 3. Linacre (2002) does not provide definitive ranges for average and expected person measures other than indicating in his example that .46

was too high, suggesting that the difference between these measures of .77 for category 1 was unacceptable.

Table 21. Rasch Analysis of Rating Scale Step 3: Summary statistics for 9 category solution

CATEGORY LABEL	OBSERVED COUNT	AVERAGE MEASURE OBS	AVERAGE MEASURE EXP	COHERENCE EXP%	COHERENCE OBS%	INFIT MNSQ	OUTFIT MNSQ	STEP CALIBRATN
1	24	-5.78	-6.49	71%	20%	2.80	2.72	NONE
2	84	-5.41	-5.29	71%	69%	.97	1.09	-7.15
3	103	-3.73	-3.64	42%	35%	.78	.75	-4.65
4	378	-2.22	-2.23	57%	57%	.98	.99	-4.21
5	484	-.82	-.77	60%	68%	.95	.95	-1.77
6	407	1.26	1.22	67%	66%	.90	.90	.35
7	366	4.11	4.19	68%	56%	.90	.85	2.71
8	670	7.49	7.42	72%	82%	.87	.87	5.28
9	335	10.01	10.05	76%	65%	1.13	1.02	9.45

Step four

As the analysis suggested that categories 3 and 4 may be the most problematic and as collapsing one category impacts on all the others, these two were first collapsed and the data analysed with an uneven 8 categorisation. However, this analysis still produced an infit statistic of 2.18 and an outfit statistic of 2.36 for category 1, indicating that this category needed to be combined with the adjacent category 2. An analysis was run with category 1 and 2 combined into one category and 3 and 4 combined into the next category, categorising the data from the VAS as follows:

Category 1: 0 to 25.

Category 2: 26 to 50.

Category 3: 51 to 62.

Category 4: 63 to 74.

Category 5: 75 to 87.

Category 6: 88 to 99.

Category 7: 100 and above entry level.

The summary statistics for ratings over the whole sample provided a person reliability of .98 and item reliability fell slightly to .96. However, this minor loss of reliability was offset by noticeable improvement in the statistical indicators regarding how well these categories function as a basis for deriving a meaningful measurement of student competency in their work placements as based on their CE's ratings of their performance.

As can be seen on Table 22, all categories meet the requirements suggested by Linacre (2002) including the stricter requirement of an outfit mean square of 1.2 or less. Category 1

does have a larger difference between observed and expected measures than other categories and a higher infit mean square statistic of 1.32 suggesting there may be some idiosyncratic usage of this category (Linacre, 1995). However, this would not be sufficient to threaten the overall measurement properties of the instrument if the VAS scores were treated as 7 categories of responses.

A 6 category solution was explored through collapsing category 1 and 2. This very slightly improved the fit statistics for the total sample (range .8 to 1.2) and brought the infit mean square statistic for category 1 down to 1.13. However this compressed the range of measurement of person abilities available on the assessment as thresholds ranged from -5.06 to 6.24 as compared to -7.94 to 7.86 for 7 categories. Consultations with expert CEs also indicated an intuitive understanding that it is possible to distinguish more than one large category of novice performance in the area of the scale below the halfway mark on the VAS. This suggests that the 6 category solution, with the first category extending to the halfway point of the VAS, would not be sufficiently discriminating in practice (M. Lincoln and A. Russell, personal communication, July 2004).

The person separation index scores further supports the 7 category division of the VAS measurements. This index indicates how many ‘groups’ of person ability appear to exist within the raw data. The index for the 34 category solution was 6.18, and rose to 7.85 for the 17 category solution and 7.88 for 9 categories, suggesting that at least 7 categories of person ability were present in the data. The person separation index did drop to 7.46 for the 7 category solution but still supports this categorisation, suggesting that this was a reasonable albeit conservative solution. Interestingly, if the data is further collapsed into 6 categories as described above, the person separation index falls to 6.87, suggesting that the degree of measurement precision is compromised. The various category options and the degree to which they meet criteria for determining rating scale categories is summarised in Table 23.

Table 22. Rasch Analysis of Rating Scale Step 4: Summary statistics for 7 category solution

CATEGORY LABEL	OBSERVED COUNT	AVERAGE MEASURE OBS	AVERAGE MEASURE EXP	COHERENCE EXP%	COHERENCE OBS%	INFIT MNSQ	OUTFIT MNSQ	STEP CALIBRATN
1	82	-7.88	-8.06	78%	63%	1.32	1.09	NONE
2	481	-4.56	-4.54	73%	67%	1.03	1.02	-7.94
3	484	-2.51	-2.45	59%	70%	.91	.92	-3.44
4	407	-.33	-.37	67%	66%	.87	.91	-1.28
5	366	2.50	2.59	68%	57%	.89	.85	1.11
6	670	5.90	5.82	73%	82%	.86	.87	3.68
7	335	8.45	8.50	77%	67%	1.12	1.02	7.86

Table 23. Summary of Category Options for VAS and the Degree to Which They Meet Guidelines for Determining Rating Scale Categories

Guideline		Category options for VAS 34	17	9	7	6
	Essential (Linacre 2003)					
1	At least 10 observations of each category	No	Yes	Yes	Yes	Yes
2	Regular observation distribution	Yes	Yes	Yes	Yes	Yes
3	Average measures advance monotonically with category	No	Yes	Yes	Yes	Yes
4	Outfit mean-squares between 0.8 and 1.2	No	No	No	Yes	Yes
	Desirable (Linacre 2003)					
5	Step calibrations advance	No	No	Yes	Yes	Yes
6	Ratings imply measures, and measures imply ratings (Coherence OBS >40%)	No	No	No	Yes > 56%	Yes >56%
7	Step difficulties advance by at least 1.4 logits	No	No	No	Yes	Yes
8	Step difficulties advance by less than 5.0 logits	Yes	Yes	Yes	Yes	Yes
	Relevant (Curtis 2004, Curtis and Boman 2004)					
9	Person Reliability	.97	.98	.98	.98	.98
10	Item Reliability	.95	.98	.98	.96	.96
11	Range of person measures (logits)	8.35	14.52	19.48	20.65	16.21
12	Person separation index	6.18	7.85	7.88	7.46	6.87

Step five

As a final confirmation for the 7 category solution, the way in which the rating categories functioned for each of the individual 11 items rated, rather than the 11 items as a group was also examined. This is also known as a Partial Credit analysis (Bond & Fox, 2001).

The statistics for each item were affected by the fact that category 1 frequently had less than 10 observations for each item. Outfits were sometimes higher than the 1.2 value

specified for the data as a whole but only two items (GC Unit 4, Professional Role; CBOS Unit 5, Planning, Maintaining and Delivering Services) had outfits greater than 2.0 for category 1. Coherence and difference in average measures were also poor for category 1 on 2 items (GC Unit 3, Lifelong Learning; GC Unit 4, Professional Role).

The difference between average measures for category 1 was also high for the majority of items (9 from 11 rated units). When items were examined for the 6 category solution, therefore increasing the number of observations due to the combination of category 1 and 2 to create one category, all categories for all items performed acceptably. However, this is of course at the sacrifice of the measurement precision provided by the 7 category solutions. The step calibrations for all of the items showed a very similar distribution pattern to the category step calibrations for the total data.

This analysis indicated that none of the items required a different categorisation to provide adequate measures for analysis. The 7 rating categories solution appeared to categorise the variable adequately for each item according to Linacre's guidelines within the constraints of insufficient observations to produce stable measures in category 1.

7.2.3. Summary: Analysis of Rating Scale Functioning

Analysis indicates that when scores from the VAS scale are divided into the 7 categories described it enables the calibration of a measurement tool with strong qualities of reliability and precision. One can have confidence that a rating in each successive category constitutes, in all probability, a clear distinction between different levels of competence in speech pathology practice. The fit statistics for this categorization are summarised in Tables 24 and 25, and Table 22 summarised how well each categorization fits the guidelines for collapsing rating scale categories.

Table 24. Rasch Analysis of Rating Scale: Summary statistics for 7 category solution for persons (n= 270)

RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN 45.6	10.5	1.22	.59	.93	-.5	.93	-.6
S.D. 18.3	1.0	4.97	.13	.81	1.6	.81	1.6
MAX. 76.0	11.0	10.30	1.10	5.61	5.3	5.18	5.4
MIN. 11.0	7.0	-10.35	.46	.07	-3.5	.06	-3.5

REAL RMSE .66	ADJ.SD 4.93	SEPARATION 7.46	PERSON RELIABILITY .98		
MODEL RMSE .60	ADJ.SD 4.94	SEPARATION 8.21	PERSON RELIABILITY .99		
S.E. OF PERSON MEAN .30					
WITH 31 EXTREME PERSONS = 301 PERSONS		MEAN 1.72	S.D. 5.76		
REAL RMSE .78	ADJ.SD 5.70	SEPARATION 7.31	PERSON RELIABILITY .98		
MODEL RMSE .74	ADJ.SD 5.71	SEPARATION 7.76	PERSON RELIABILITY .98		

Table 25. Rasch Analysis of Rating Scale Step 4: Summary statistics for 7 category solution for items (n=301)

RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN 1119.9	256.8	.00	.11	.95	-.6	.94	-.8
S.D. 79.1	13.6	.60	.00	.15	1.6	.13	1.4
MAX. 1229.0	268.0	1.07	.12	1.22	2.3	1.17	1.5
MIN. 976.0	224.0	-1.07	.11	.75	-3.0	.76	-2.7

REAL RMSE .12	ADJ.SD .59	SEPARATION 5.12	ITEM RELIABILITY .96		
MODEL RMSE .11	ADJ.SD .59	SEPARATION 5.23	ITEM RELIABILITY .96		
S.E. OF ITEM MEAN .19					

7.3. Calibrating the Assessment Tool

7.3.1. Process

7.3.1.1. Introduction

Every measurement tool requires careful calibration to ensure that the measurements it makes are as accurate as possible. Rasch analysis provides calibrations for items such as standard errors and fit statistics, and for the category step calibrations or thresholds. Recent work by Curtis (2004) and Curtis & Boman (2004) has identified that inclusion of data from persons whose response pattern does not fit the pattern predicted by the Rasch model decreases the precision of item parameter estimates and truncates the range of threshold estimates further affecting the measurement precision of the estimates. Removing persons identified as having poor fit with the model for the purpose of calibrating the assessment tool promotes better precision of the tool as they are introducing variance that is unrelated to the underlying trait being assessed.

Curtis and Boman (Curtis, 2004; Curtis & Boman, 2004) identified that the infit mean square (IMS) measure for persons was most critical in this process due to its sensitivity to deviations from expectation for well targeted items and persons. Thus if the person's ability level was well targeted, as indicated by the rating pattern they received on the items (See Section 6.3.3. for an explanation of person fit statistics), the infit mean square (IMS) would be within an acceptable range. They found that the person IMS needs to be at least 1.55 or above before excluding a person's data as misfitting and suggested that excluding persons with an infit mean square of 1.8 or above minimised the risk of losing relevant information for measurement while accepting a reasonable level of 'noise' in the data (Curtis, 2004; Curtis & Boman, 2004).

There is also a case for removing overfitting persons, as indicated by low IMS, but Curtis and Boman found it is primarily the high IMS statistics that most affect the precision of the measurement tool. Given the assessment tool required ratings by judges, which are more likely to produce overfitting persons (Linacre, 1998a) and therefore low IMS, it was

conservatively decided to retain all overfitting persons when calibrating the tool and only exclude those with Infits of 1.8 and above[9].

7.3.1.2. Examining Misfits

Thus the criteria for removal of misfitting persons, for the purpose of calibrating the tool prior to generating a measure of competence for the students assessed by the research tool, was conservatively set at any person with an IMS score of 1.8 or greater. Before proceeding with calibrating the assessment instrument without misfitting data it is recommend that the ratings be examined to ensure that there are no weaknesses in the tool creating these anomalies (Curtis, 2004). This enables further analysis to proceed on the assumption that students with high IMS scores are being accurately identified by the assessment tool on the basis of inconsistencies in their development of competency.

Twenty-two students met this criterion (Table 27) and (as would be expected by the high IMS) show very variable patterns of ratings, some spanning at least 3 categories, suggesting that their ability level was not well targeted. Given that these students are introducing unacceptable amounts of variance into the data, suggesting that perhaps the assessment tool does not succeed in determining the level of competence that they possess, it is important to examine their ratings and any information that may account for this variability. There were a number of possibilities that could explain the variable ratings.

First, 6 of the 22 students failed their end placement assessment (students 2, 4, 31, 100, 101, 121). It is plausible that they in fact failed due to such variable performance and clearly their CEs had assessed their overall performance as being below the level expected for that particular placement. Variability is considered to be a common characteristic of students who are marginal or struggling to meet the performance criteria in a work placement (Robertson et al., 1997) and this is likely to be reflected in variable ratings from their CEs.

A further 6 students (51, 92, 214, 218, 290, 295) were very early in their clinical development, having accrued 84 hours or less of experience, yet showed a number of unexpectedly high scores frequently in relation to Generic Competencies (the first 4 ratings on the table). These students may in fact be showing an uneven pattern of development due to performing above expectations in more generic skill areas but rating at lower levels on more specific competencies due to their lack of experience in clinical work.

[9] The sub sample of the data that is comprised of all those students whose scores have IMS values below 1.8 will be referred to as the Calibration Sample.

This leaves 10 students whose pattern is not so easily accounted for on the basis of related information collected through the research. There are of course many potential sources of variation including idiosyncratic use of the scales by CEs, students demonstrating expected baseline ratings for their level of experience but having some specific personal knowledges or experiences leading to higher ratings for these competencies, or conversely demonstrating specific weaknesses.

Relatively higher ratings were sometimes observed (5 out of the 10 cases) amongst the Generic Competencies (the first 4 ratings in the string in Table 27). It is possible that some students bring high levels of personal and professional skills to their work as student speech pathologists and require further experience to develop competency in the occupationally related skills of their chosen profession. Finally, lack of opportunity to practice and develop a particular competency due to the experiences provided at the placement site may result in an unexpectedly low rating for a particular competency area. Overall, however, there does not appear to be a specific weakness of the assessment instrument, e.g. a particular item being rated variably, that would suggest that the construct of unidimensionality is contravened.

Table 26. Students With IMS Values >1.8 in the Total Sample

Person No.	IMS	Rating patterns for 11 items (* denotes missing rating)
2	4.92	46377547566
4	2.50	21211132321
31	4.06	32214332142
34	2.99	22522123232
37	2.23	55653245435
48	2.27	4442**354*5
51	3.29	224534241*2
78	2.91	22313223211
80	2.30	76676666646
92	2.19	34252*24222
100	4.96	43525552532
101	5.61	11142124313
107	3.76	56675455777
121	2.56	23245243244
128	1.95	55675556635
131	3.13	55775556366
173	2.42	44435434413
214	2.24	4632244*4**
218	3.88	27343323332
245	1.84	66675576767
290	2.51	35453222322
295	2.67	42424322222

7.3.2. Results

Once the 22 misfitting cases were removed the Calibration Sample data was entered into Bigsteps (Linacre & Wright, 2003) and coded into the 7 categories previously described. As predicted, removing the misfitting cases resulted in an expanded range of measures as demonstrated by the logit values for the step calibrations now ranging from –10.93 to 10.11 (Table 28). The difference between the expected and observed average measures and coherence measures, although already acceptable before misfitting persons were removed,

was also smaller for each category in the Calibration Sample. Both item and person reliabilities increased by .01 to .97 and .99 respectively, indicating a higher degree of measurement precision.

A logit step of 6.40 now existed between categories 1 and 2 and a step of 5.01 between categories 6 and 7, both exceeding Linacre's suggested step value of no greater than 5 logits (Linacre, 2002). However, this recommendation was not made in the context of more recent work on removing misfitting persons from the data for calibration purposes and analysing the data from Calibration Sample (Curtis, 2004; Curtis & Boman, 2004). In addition, the recommendation falls into the 'non essential' category and is to be considered along with other information on the measurement precision of the scale categories – all of which are indicating improved and excellent measurement properties.

All data was then re-entered into Bigsteps with the 7 thresholds anchored at the values identified by the calibration analysis. Anchoring involves specifying the threshold values for the rating scale categories for the assessment items to be used by the program for all data subsequently analysed. These values are set at the levels identified as providing the highest level of measurement precision as identified by the calibration procedure. These step calibrations enable the category into which a person measures to be determined. For example, ratings that generate a person measure of -12.0 indicate that the students' level of competency falls into category one, a measure of 7.0 logits would suggest that the students' level of competency falls into category 6.

Table 27. Summary Statistics for 7 Categories for the Calibration Sample

CATEGORY LABEL	OBSERVED COUNT	AVERAGE MEASURE OBS	AVERAGE MEASURE EXP	COHERENCE EXP%	COHERENCE OBS%	INFIT MNSQ	OUTFIT MNSQ	STEP CALIBRATN
1	64	-11.50	-11.6	78%	75%	1.18	.89	NONE
2	423	-6.18	-6.16	78%	73%	1.01	.99	-10.93
3	446	-3.16	-3.12	65%	73%	.90	.89	-4.53
4	368	-.12	-.14	71%	69%	.94	.94	-1.50
5	328	3.63	3.62	69%	59%	.89	.88	1.75
6	644	7.77	7.76	75%	82%	.96	.95	5.10
7	318	10.91	10.91	76%	70%	.92	.90	10.11

7.4. Examining Items

7.4.1. Items Rated

7.4.1.1. Results

The end assessment units were rated for the majority of students (Table 29). Rasch analysis is able to manage missing data without compromising the accuracy of the analysis (Bond & Fox, 2001). A rating is coded as missing if CEs indicate that it was ‘not observed’ or simply does not provide a rating for that item (which occurred with some hard copy data, but was prevented in the online data). It can be seen from this table that students were least likely to receive sufficient experience for CEs to rate on items 5 and 6, the CBOS Units for Assessment and the related activity of Analysis and Interpretation.

Interestingly mid placement data suggested that fewer students were involved at mid placement in observable activities for some competencies with more than 30 data points missing for element level ratings of Items 5 and 6, and also 7, 9 and 10 (Table 30). These figures were lower than for other Units and some were particularly low suggesting that judgements of competency are being made on fewer observations for these items if ratings are given.

Table 28. Number of Ratings Provided for Each Item

Item Number	Competency Name	Number of ratings (N=299)
1	GC Unit 1, Clinical Reasoning	290
2	GC Unit 2, Professional Communication	288
3	GC Unit 3, Lifelong Learning	291
4	GC Unit 4, Professional Behaviour	290
5	CBOS Unit 1, Assessment	251
6	CBOS Unit 2, Analysis and Interpretation	262
7	CBOS Unit 3, Planning of Speech Pathology Intervention	299
8	CBOS Unit 4, Speech Pathology Intervention	299
9	CBOS Unit 5, Planning, Maintaining, Delivering Speech Pathology Services	299
10	CBOS Unit 6, Professional, Group and Community Education	283
11	CBOS Unit 7, Professional Development	298

Table 29. Mid Placement Ratings Representing More Than 30 Data Points Missing From a Possible 311

Element of competency rated	N of ratings (N=311)
CBOS 1 Assessment	
CBOS1.1 Establishes and documents the presenting communication and/or swallowing condition and issues; identifies the significant other people in the client's life and collates information on the client.	241
CBOS1.2 Identifies the communication and/or swallowing conditions requiring investigation and the most suitable manner in which to do this.	247
CBOS1.3 Administers speech pathology assessment relevant to the communication and/or swallowing information required.	244
CBOS1.4 Undertakes assessment within the ethical guidelines of the professional and all relevant legislation and legal constraints, including medico-legal responsibilities.	239
CBOS 2 Analysis and Interpretation	
CBOS2.1 Analyses and interprets speech pathology assessment data.	246

CBOS2.2 Identifies gaps in information required to understand the client's communication and swallowing issues and seeks information to fill those gaps.	252
CBOS2.3 Determines the basis or diagnosis of the communication and/or swallowing issues or condition and projects the possible outcomes.	216
CBOS2.4 Reports on analysis and interpretation.	178
CBOS2.5 Provides feedback on results of interpreted speech pathology assessments to the client and/or significant others and referral sources, and discusses management.	185
CBOS 3 Planning of Speech Pathology Intervention	
CBOS3.3 Discusses long-term outcomes and decides, in consultation with client, whether or not speech pathology strategies are appropriate and/or required.	228
CBOS3.4 Selects speech pathology program or intervention in conjunction with the client and significant others.	255
CBOS3.5 Defines roles and responsibilities for the management of the client's swallowing and/or communication condition and issues.	285
CBOS3.6 Documents speech pathology intervention, plans, goals, outcomes, decisions and discharge	270
CBOS 5 Planning, Maintaining, Delivering Speech Pathology Services	
CBOS5.3 Uses service provider's electronic systems.	226
CBOS5.5 Updates, acquires and/or develops resources.	280
CBOS5.6 Consults and coordinates with professional groups and services.	208
CBOS5.8 Collaborates in research initiated and/or supported by others.	107
CBOS5.9 Participates in evaluation of speech pathology services.	60
CBOS Unit 6, Professional, Group and Community Education	
CBOS6.1 Identifies the practice of speech pathology in a range of community contexts.	190
CBOS6.3 Undertakes preventative, educational and or promotional projects or programs on speech pathology and other related topics as part of a team with other professionals.	88
CBOS6.4 Demonstrates an understanding of principles and practices of clinical education.	256

CBOS Unit 7, Professional Development	
CBOS7.3 Undertakes preventative, educational and or promotional projects or programs on speech pathology and other related topics as part of a team with other professionals.	275

7.4.2. Unidimensionality

As described in Sections 6.3.1. and 6.3.2., the Rasch analysis model provides information as to whether all the items sample the same underlying trait or at least a set of underlying personal factors that function in unison to determine the students' performance in the same way on each item (Bond & Fox, 2001). This is described in Rasch terms as "unidimensionality" (Bond & Fox, 2001) and is central to the validity of an assessment tool.

7.4.2.1. Process

Fit statistics are critical for examining whether an item is contributing useful information about the variable under examination. If an item does not produce ratings that fit the expected pattern it is assumed that the fault lies with the item e.g. it does not sample the same construct as other items, it is not well written, or the rating scale is interpreted differently by raters than intended by the test designer and so produced unexpected responses. Bond and Fox (2001) suggest that, for a high stakes test, a range of 0.8 to 1.2 mean squares values is reasonable. However, they also recommend that, for a situation where judgement agreement is encouraged (e.g. through training), some overfitting of the model may be permissible i.e. ratings being more regular than predicted such that fit statistics as low as 0.4 could also be acceptable.

7.4.2.2. Results

Fit statistics for items as assessed by the Calibration Sample (misfitting persons removed) are all within this range the lowest being .81 and the highest 1.17 (Table 31). This confirms that each of the items contributes usefully to measurement of the construct of competence thus the assumption that the assessment tool is assessing a unidimensional trait of 'competence' is strongly upheld (Bond & Fox, 2001). This represents the first major finding of this research, namely that the assessment tool is indeed measuring what it was designed to measure – a unidimensional construct of workplace competency of speech pathology students.

Table 30. Item Statistics for Calibration Sample

ENTRY NUMBR	RAW SCORE	COUNT	MEASURE	ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTBIS CORR.	ITEMS
1	1016	240	.50	.13	.85	-1.7	.84	-1.6	.96	I0001
2	1039	237	-.22	.13	.88	-1.3	.84	-1.5	.96	I0002
3	1060	239	-.37	.13	.91	-.9	.92	-.7	.95	I0003
4	1109	238	-1.29	.13	1.01	.1	1.01	.1	.95	I0004
5	898	203	.88	.14	1.00	.0	1.00	.0	.95	I0005
6	911	216	1.32	.14	.94	-.7	.85	-1.4	.95	I0006
7	1066	246	.24	.13	.81	-2.1	.82	-1.7	.96	I0007
8	1141	246	-1.04	.13	.83	-2.0	.81	-1.7	.96	I0008
9	1096	246	-.26	.13	.98	-.3	.98	-.2	.95	I0009
10	1019	235	.55	.13	1.17	1.7	1.08	.7	.94	I0010
11	1095	245	-.30	.13	1.02	.2	1.00	.0	.95	I0011
MEAN	1041.	236.	.00	.13	.94	-.6	.92	-.7		
S.D.	74.	13.	.75	.00	.10	1.1	.09	.8		

7.4.3. Item Reliability

7.4.3.1. Results

As mentioned above, the assessment has excellent Item Reliability under all conditions. Item Reliability was measured at .97 when all data was entered and anchored to the thresholds determined by the Calibration Sample. This indicates that these items can be expected to maintain the same estimates of item difficulty if used to rate a group of students with similar ability levels.

7.4.4. Differential Item Functioning

7.4.4.1. Process

Examining the quality of a measurement or assessment tool also requires determining whether the items have significantly different meanings for different groups within the sample, termed differential item functioning analysis or DIF (Bond & Fox, 2001). Traditionally such analyses have been used to examine if a particular subgroup, e.g. girls versus boys, are disadvantaged by the way a particular item is written e.g. on a mathematics test. If an item(s) does not function the same for a particular subgroup, e.g. those who had their assessment submitted online vs. hard copy, this indicates that the rating students receive on this item may be affected by this factor rather than the underlying trait of competence they are judged as possessing.

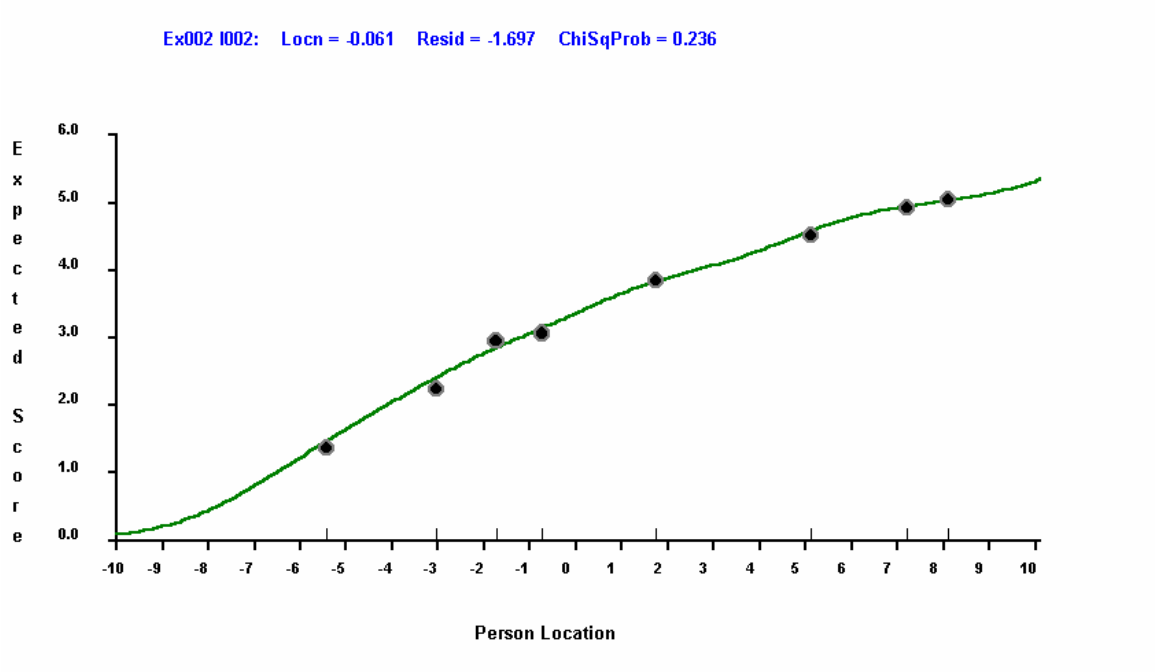
The focus therefore is on assessing how consistent item parameters are across different subgroups. It can be expected that factors such as the timing and sequence of placements and teaching across different programs will affect the level of competence a particular group may achieve at a particular point of time on a particular item – but this is not a function of a poorly written or interpreted item. Some of these differences are explored in the subsequent section on person measures and include university program attended and CEs' level of experience.

The primary factor that may affect student ratings of the assessment tool, that is not sample dependent and can reasonably be expected to be influential upon student performance or CEs' ability to rate accurately, is that of the format in which the assessment tool was used: online or hard copy. It was essential to determine early in the analysis whether the data submitted by hard copy was significantly different to that submitted online. If there were differences subsequent analysis would need to be conducted on each of these groups separately and the hypothesis that the manner of submission would not significantly affect the way in which students were rated would be proven to be false. DIF analysis determines whether the subgroups of online versus hard copy behave in similar ways on the items such that the items can be considered to have the same relative difficulties for all persons in the sample regardless of the format by which they are presented.

The data was entered into RUMM 2020 (Andrich & Sheridan, 2003) and the items were examined for differential functioning according to the online verus hard copy subgroups. RUMM produces statistics and graphs to assist in this calculation. First, Item Characteristic Curves (ICC) are produced for each item. Fig. 14 is an example of an ICC calculated for Item 1 (GC Unit 2, Professional Communication) using 7 categories, with original data, and no thresholds anchored.

The ICC is the expected value curve expected by the Rasch model for every possible person ability. The measured ability for each person is derived by their responses to all the items on the assessment tool, taking into account the difficulty of each item, and transformed into an interval logit scale. The formula for the Rasch modelling graphs an ICC that represents the score each level of ability would be predicted to receive as a function of the particular level of difficulty for that item.

Figure 14. Example of an Item Characteristic Curve (ICC)



On Fig. 14 it can be seen that as the person ability levels increase, as indicated by their measured ability location on the logit scale (X axis), the predicted score on this item increases, as indicated by 'expected score' on the Y axis. Note that RUMM recodes ratings of 1 as 0, 2 as 1, 3 as 2 and so forth. Thus a person with a measured ability of –.8 has an expected score of 3 representing a rating of four [10].

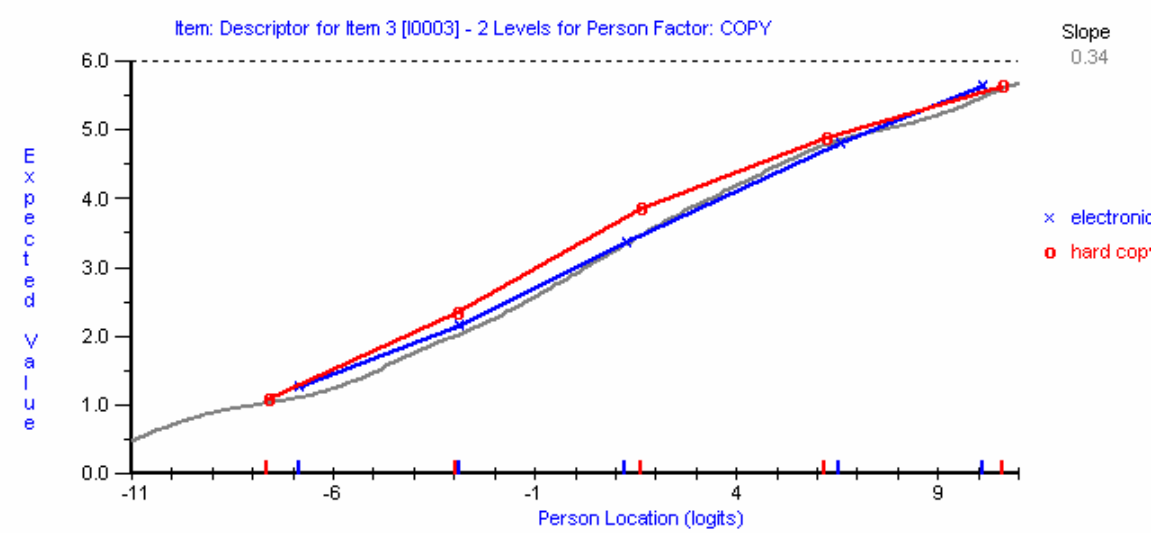
As there are not enough people in the sample who have the same total ability level for the entire ability range to identify whether the relationship observed between ability and ratings is sufficiently similar to the one predicted by the ICC curve, class intervals are used to plot whether the ratings observed are similar to the ones predicted. The 10 dots represent 10 Class Intervals (CI) which are 10 approximately equal groupings of people with similar ability levels as assessed by the whole assessment, and each dot represents the mean ability level for each group. These are plotted against the ICC to see whether their positions closely correspond with the predicted ICC.

When examining DIF a similar process is used with the addition of each class interval being divided according to the sample characteristics being examined. For example, the sample can be divided into those who have their assessment submitted by hard copy and those submitted online. Each CI will have a mean score plotted separately for these two characteristics, these are graphed, and their similarity to the modelled expected score for the

[10] The ability levels in this figure are produced from unanchored analysis and so vary from the ability levels in the anchored analysis used for examining the DIF of the items.

item (ICC) is examined to see if it is statistically different through the use of an ANOVA. Figure 15 is an example of a DIF graph for online vs. hard copy submission for Item 2 (GC Unit 2, Professional Communication). The grey line represents the modelled or ICC probability and the coloured lines marked by x and o represent the two different groups under examination. RUMM provides ANOVA information on how different these groups are from each other (online and hard copy).

Figure 15. Example of a Differential Item Functioning graph



From the graph it can be seen that persons whose ability level is estimated from data yielded by online copy (n = 236) more closely approximates the ICC than persons assessed by hard copy data (n=85). However, the ANOVA comparing the online and hard copy observations is not significant at the .05 level (p = .465) suggesting that the ratings each of these groups assign to students of particular overall ability estimates are more similar than different to each other.

The DIF analysis for the hard versus online copies included those persons who had high IMS in the original analysis. This has the advantage of using a higher N and therefore giving the DIF testing more power. However it does mean that persons who have scores with a wider range of standard errors are included, creating greater variance within each of the comparison groups and thus reducing the power of the DIF test and possibly identifying fewer items as functioning significantly differently for the groups being compared (D. D. Curtis, personal communication, October 2004). In balance, it was decided to use this sample rather than the Calibration Sample as it models the real life scenario of the assessment tool being applied to all students.

7.4.4.2. Results

Table 32 provides the ANOVAs of the estimated person abilities for the Hard vs. Online submission groupings for each item. Data from 85 hard copy submission and 236 online submissions were suitable for analysis.

Table 31. Differential Item Functioning ANOVAs Between Ratings Submitted Via Hard Copy or Online Submission

Item Number	Competency Name	ANOVA: Hard vs. Online
1	GC Unit 1, Clinical Reasoning	0.059
2	GC Unit 2, Professional Communication	0.077
3	GC Unit 3, Lifelong Learning	0.465
4	GC Unit 4, Professional Behaviour	0.071
5	CBOS Unit 1, Assessment	0.637
6	CBOS Unit 2, Analysis and Interpretation	0.185
7	CBOS Unit 3, Planning of Speech Pathology Intervention	0.820
8	CBOS Unit 4, Speech Pathology Intervention	0.412
9	CBOS Unit 5, Planning, Maintaining, Delivering Speech Pathology Services	0.907
10	CBOS Unit 6, Professional, Group and Community Education	0.417
11	CBOS Unit 7, Professional Development	0.123

As can be seen from this table, there were no significant differences at a .05 level between the expected data and data submitted by online or hard copy. This analysis indicates that the data generated by both types of systems can be combined for future analysis.

However, the graphs (Appendix 25) do indicate that there was some variability in how the different versions were used in relation to different ability levels illustrating why some items were approaching significance at the .05 level. The DIF graphs for most items suggest that CI 2 and 3 were either more likely to receive lower ratings (4 items) or higher ratings (5 items) on hard copies than predicted by the ICC. Ratings generally converged at the higher and more critical CIs of above 6 logits i.e. those that approached the decision making point of whether students were at entry-level.

The majority of online data was submitted by CEs employed at The University of Sydney (161 of 231 assessments submitted online). This may have created some sample dependent effect upon the DIF as a result of specific differences in student performances owing to the particular teaching and practicum program provided by The University of Sydney. In addition, a rater community effect where there is a strong shared interpretation of the assessment items could be operating due to 152 of these assessments being submitted by CEs employed by The University of Sydney. Given these possible sources of influences, the lack of any significant difference between the online and hard copies suggests that the rating procedure is very robust.

7.4.5. Item Difficulty

7.4.5.1. Process

Rasch analysis provides information on the difficulty level of each item for which the students are rated. This identifies which items are the most difficult to be rated highly on and vice versa. Table 33 provides the estimation for each item difficulty (and error range) for the Calibration Sample in terms of the interval measure generated by Rasch analysis (logits) which is derived from the pattern of ratings for each student assessed on the tool.

7.4.5.2. Results

These measures indicate that CEs found it hardest to rate students of any given ability level highly on Item 6, CBOS Unit 2, Analysis and Interpretation and easier to give a high rating for Item 4, GC Unit 4, Professional Behaviour (Table 33). Thus the students were most likely receiving a rating as ‘entry level competent’ of 7 for Item 4 before they received a rating of 7 for item 6. However, as can be seen from Table 33, the item difficulties represent only a small logit range of 2.61 (-1.29 to 1.32) when compared to the person measure range and suggest that the differences between item difficulties are relatively small.

Table 32. Rasch Measurement of Difficulties of Assessment Items

Item Number	Competency Name	Order of Difficulty	Rasch measure (logits)	Error (logits)
6	CBOS Unit 2, Analysis and Interpretation	1	1.32	.14
5	CBOS Unit 1, Assessment	2	.88	.14
10	CBOS Unit 6, Professional, Group and Community Education	3	.55	.13
1	GC Unit 1, Clinical Reasoning	4	.50	.13
7	CBOS Unit 3, Planning of Speech Pathology Intervention	5	.24	.13
2	GC Unit 2, Professional Communication	6	-.22	.13
9	CBOS Unit 5, Planning, Maintaining, Delivering Speech Pathology Services	7	-.26	.13
11	CBOS Unit 7, Professional Development	8	-.30	.13
3	GC Unit 3, Lifelong Learning	9	-.37	.13
8	CBOS Unit 4, Speech Pathology Intervention	10	-1.04	.13
4	GC Unit 4, Professional Behaviour	11	-1.29	.13

7.4.7. Summary

The analysis strongly supports the premise that the assessment tool is composed of items that consistently support CEs' judgement and assess a unidimensional or coherent concept of competency. The format of the assessment tool (hard copy or online) does not affect the way in which the assessment items are used. Once the appropriateness of the assessment items has been determined, its effectiveness in measuring the competency of students can be evaluated.

7.5. Examining Persons

7.5.1. Process

Rasch analysis provides several statistics that give information on person ability. A Rasch score is generated for each person and is an interval measure (logit) that quantifies the amount of competency each person is determined to have on the basis of the sum of their scores on

the 11 items on the rating scale. The accuracy of this measure is qualified by a standard error being calculated for each person. Fit statistics are also determined for each person and identify how closely the rating pattern they have received conforms to the Rasch measurement model. In addition the person measure is an interval measure that conforms to the assumptions required for classical statistical analysis so these methods can be used to examine the probability of various relationships within the data. The effectiveness of the assessment tool in measuring the competency of students is evaluated by examining each of these statistics.

7.5.2. Person Reliability

7.5.2.1. Results

The Person Reliability score of .98 indicates that the ability estimates of the person assessed by this tool are well targeted by the items and the ordering of person ability as measured by the Rasch score would be highly likely if a similar set of items that effectively identified competency were administered. Thus a large spread of ability and a clear hierarchy of ability or development on the construct are identified by the assessment. This is confirmed by the very wide spread of person ability from the minimum measure of -14.24 through to the maximum measure of 13.41, a 27.65 logit range. Ranges as small as 6 logits have been identified in the literature as satisfactory (Linacre, 2002) as are person reliabilities of above .80 (Curtis & Denton, 2002).

7.5.3. Determining Fit Statistics Range

7.5.3.1. Background

The previous section on calibrating the assessment tool mentioned the importance of IMS values for determining whether the persons measured are actually being measured accurately by the assessment tool. The Rasch model presumes that there will be some variability in the scores that persons receive on each item on which they are rated. The fit statistics summarise the difference between what is observed in the data and what was expected (Linacre, 2001). These statistics are estimated in two ways. Outfit mean squares are the unweighted estimates of the degree of fit of response and tend to be influenced by off target observations (Bond & Fox, 2001). Infit mean squares (IMS) are weighted to give more value to on target observations and are more sensitive to irregular inlying patterns (Bond & Fox, 2001). It is the

IMS that identify rating patterns that are most relevant for identifying those persons that fit the Rasch model (Curtis, 2004; Curtis & Boman, 2004).

In the simplest Rasch model, where dichotomous data is generated through scoring items as right/wrong, it is expected that as the items get harder and harder only persons of higher ability will get them right. However, it is also expected that this progression will not be perfect, e.g. the first 10 items right and the next 10 answered incorrectly. This is termed the Guttman pattern where 1 = correct and 0 = incorrect and would look like this: 11111111110000000000. This data would be considered to be overfitting the Rasch model or too perfect. The model presumes there will be some variation around the point at which the person's ability actually lies i.e. this is modelled as being the point at which there is a 50/50 chance of getting the item correct. Thus the modelled progression over 20 items is more likely to be: 11111111011001000000. If the pattern of scores is not similar to this, then other issues may be influencing the person's score (see Table 34 for suggested explanations for various patterns). The fit statistic value that is considered to match the variation in ratings expected is set at 1.0, rather than a value of 0 which would indicate that there is absolutely no variation between the predicted score or rating, and the score observed.

When this model is extended to rating scale (polytomous) data it is conceptualised slightly differently. In this case it is assumed that students will not receive exactly the same rating for every item, this would produce an overfitting case with a low IMS, but also that the ratings will not vary too widely producing an underfitting case with a high IMS. Table 34 gives some examples from the research data of what the rating strings look like for each of these cases as well as for a well fitting person.

It can be seen from this table that the model suggests that some variation in the score in the fitting cases can be considered as still producing a valid Rasch score but that the wide variation or unexpectedly low/high scores may cause us to think that the Rasch score is not truly representative of the students' overall ability. For example, the underfitting string for the Rasch score of 7.61 suggests that it might be possible that this student may have been measured as having a higher level of ability than the fitting string for the same score, if it wasn't for an unexpectedly low rating on the 10^{th} item meaning that the person's ratings spread over 4 categories on the 7 category rating scale.

Table 33. Examples From the Research Data of Rating Strings and Related IMS Values Illustrating Degree of Fit to the Rasch Model

Fit	Rating String	IMS value	Rasch Score
Underfit	27343323332	5.23	-2.46
	76676666646	3.05	7.61
Fitting	34442334233	1.12	-2.46
	65666667666	1.06	7.61
Overfit	33433333433	.54	-2.46
	66666666666	.09	7.61

Overfitting scores are a little more difficult to interpret in the context of this assessment as rating tasks, particularly those that encourage agreement, tend to produce more overfitting data (Bond & Fox, 2001). Generally they are considered to be 'too good to be true' and it is certainly possible that overfitting strings could be produced by a rater who is not truly attending to the assessment task and so is marking students at the same place on all the scales. On the other hand, an overfitting string of ratings could be given to a student who is consistently performing at a particular level.

Smith (1996) identifies that the purpose of fit statistics is to assist in quality control of measurement through identifying data that does not match the requirement of the Rasch model, and highlights that the use of fit statistics with polytomous data is a much more recent phenomenon. Data that does not fit the model is not automatically rejected but examined to determine how and why they do not fit and the effect that this misfit may have on the measurement task (Smith, 1996).

There are therefore several issues to be addressed. First, what range of IMS values should be applied to this data to determine whether some data does not fit the model? The second issue is how to then interpret the measurement scores given to those persons whose IMS values do fall outside the acceptable range suggesting that their data does not fit the model. Finally, some determination needs to be made as to how this affects the measurement of competency.

7.5.3.2. Process

As discussed during the section on calibrating the assessment tool, Curtis and Boman (2004) identified that the critical IMS value for rejecting underfitting data without risking rejecting fitting cases was around 1.6 to 1.8, and 1.8 was decided on for calibration purposes. The appropriate value for rejecting overfitting data was suggested to be between .2 and .55 depending on the number of decision points in the data but lower IMS appeared to have less effect on the measurement qualities of the items. Programs such as Bigsteps (Linacre & Wright, 2003) use an IMS of 2.0 as the point at which persons are identified as misfitting, although no rationale is offered for this in the manual. Curtis (personal communication, October 2004) suggests that an IMS values from 1.8 to 2.2 could be defensible.

Given the way in which student performances are scored and the context of the assessment, it is proposed that a number of factors could influence the IMS values for persons. Rater behaviour is a likely factor and visual inspection of the data suggests that there were CEs who tended to rate either in an underfitting manner, where they tended to rate over a large spread of the VAS, or an overfitting manner. This research was not designed in a way that allows rater behaviour to be teased out as a factor in the scores calculated for persons and this would clearly be a useful focus for future investigations.

Some students may score lower on particular competencies due to lack of opportunities to practice a specific competency because of the opportunistic nature of the learning experiences in practicum placements, resulting in an underfitting pattern or high IMS value. For example, the underfitting student in Table 34 above who received a score of 7.61 had a much lower rating of 4 for item 10 relative to the other items. This item has been identified in midplacement data as one that students are less likely to have an opportunity to be rated on. As competency develops with experience, it is possible that a lower rating could be received for this reason. If this has occurred for this student, his/her true competency level may in fact be higher than his/her score suggests. Alternatively, students may have a specific strength or weakness in their performance profile, analogous to the student who unexpectedly gets a correct answer on a harder item on a test due to his/her specialist knowledge (Bond & Fox, 2001) resulting in a higher rating on a specific competency(s), thus a high IMS.

IMS values could also be affected by marginal performances. It is anticipated that marginal students could show two types of IMS values: underfitting or fitting but allied with a lower than required person score. Underfitting may occur because marginal students are

frequently characterised in the literature as being very variable in their performance due to their difficulty in consistently integrating all the elements required for a satisfactory performance (Robertson et al., 1997). A variable performance will produce high IMS values. It is also plausible that some marginal students are consistent in their performance but that their overall performance does not reach the level expected to pass the placement. Identifying fitting but underperforming students would require benchmarking the expected minimum score for each placement in the sequence for each university program, something this research was not designed to do.

7.5.3.3. Results

Distribution of IMS Values

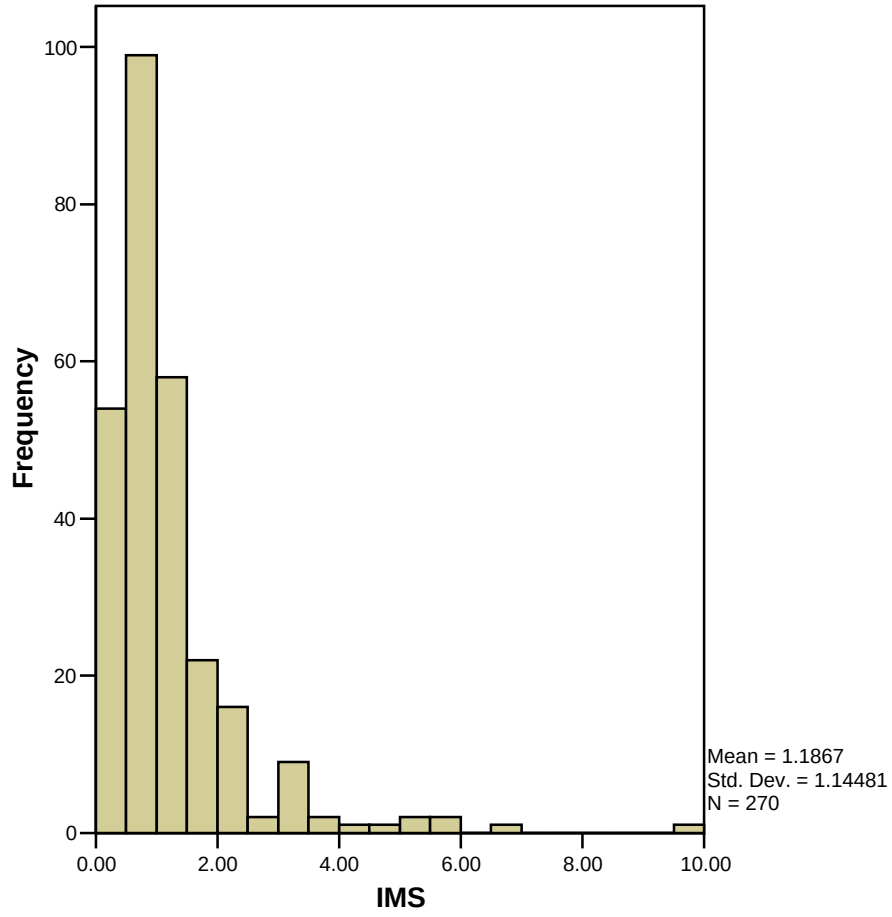
Given the design of the current research it is not possible to address or investigate all of the above propositions with a view to determining appropriate interpretation of IMS values for each student who is 'measured' by the assessment tool. However, examining how the IMS values are distributed in the data does provide information to guide the decision regarding the ranges to use. IMS values can only be calculated for scores that are not either the minimum or maximum on the assessment and for the total sample the total number of students with an IMS score for their end assessment is 270. The IMS values under consideration here are those generated after the total sample is entered into the anchored thresholds as determined by the Calibration Sample. This procedure results in a greater number of high IMS scores than thresholds determined by the Calibration Sample. This occurs because of the greater measurement precision provided by the anchored thresholds which are based only on scores in which a great deal of confidence can be held regarding the accuracy of their measurement (unlike scores with high IMS values). Thus applying a stricter measurement requirement upon the whole data set will result in more scores being identified as having doubtful measurement properties (Curtis, 2004; Curtis & Boman, 2004).

As can be seen in Fig 16, a large number of IMS values fall below 1.0, with a modal value of .88 for the whole sample. This indicates that the data tends to be overfitting compared to the ideal IMS value of 1.0. The majority of the data that yielded IMS measures[11], representing 270 assessments, falls at or below an IMS of 1.8 (84.8%) with 234 persons (86.7%) at or below 2.0 and 245 persons (89.6%) falling at or below 2.2. Thus, somewhere between 42 and

[11] Maximum or minimum scores do not have an IMS value calculated.

25 persons will be identified as underfitting, depending on where the IMS cut off score is determined, with 37 having IMS values above 2.0.

Figure 16. Distribution of IMS values for whole sample of field trial data



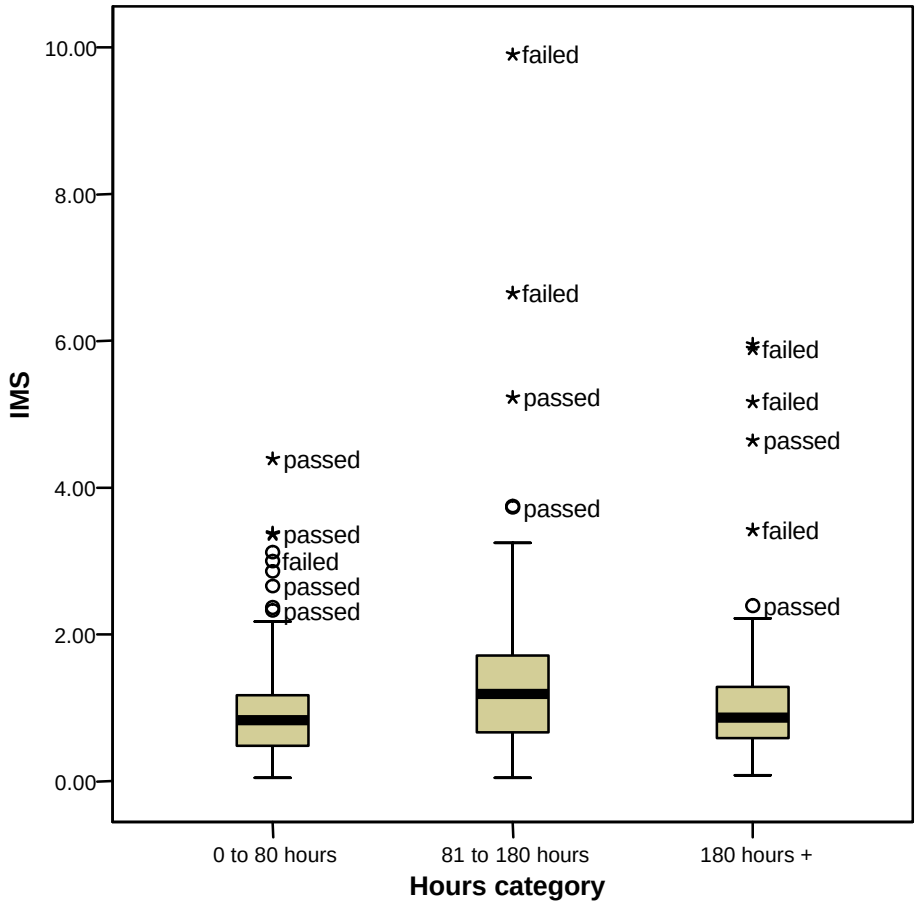
Distribution of IMS and Sample Subgroups

All subgroups within the total sample were investigated to determine if the IMS value was influenced by any particular grouping of raters. There were no significant differences found for an ANOVA comparing the IMS values generated by ratings from CEs grouped according to 3 different levels of self rated experience ($p = .081$). Students rated by CEs employed by The University of Sydney had significantly lower IMS values ($p = .000$) which further supports the notion that these CEs form a rater community with a shared understanding of competence and how it develops, and so are less variable in their judgements.

A significant difference ($p = .017$) also exists when an ANOVA is calculated to compare the 3 groupings of student experience. As can be seen in Fig 17, students in the middle group of experience were the most variable. This suggests that beginning students tended to perform

at similar levels across all competencies and that middle students were showing greater variability as they increased their skills. More experienced students had relatively less variability in their ratings than the middle group suggesting more consistency in their performance – clearly a finding that would be hoped for in terms of their development as practitioners.

Figure 17. Range of IMS values for each category of student experience



Marginal Students

Twelve sets of ratings and their associated Rasch scores and IMS values were collected on students who were identified as at risk of failing their end placement assessment. These are summarised in Table 35 where it can be seen that 10 of the 12 students had scores with IMS values of 2.17 or above, accounting for 10 of the 37 person with IMS values above 2.0. The fact that more students are identified as misfitting using thresholds anchored according to the values determined by the Calibration Sample, 10 as opposed to 6 before calibrating, suggests that the measurement precision of the assessment tool has been enhanced by this procedure.

Of the two marginal students with low IMS values, the student with the Rasch score of -11.57 had a very low rating on the VAS and did not have much variation in his/her rating string, suggesting that he/she were marginal due to a performance below the expected level for this placement. These IMS scores suggest that 2.0 or above may be the point at which the Rasch score allocated to a person requires careful consideration to determine whether it is accurate and how it should be interpreted.

The usefulness of the IMS values in identifying students with highly variable performances and the effect it may have on their overall measure is indicated by one student who had two supervisors concurrently. Both supervisors indicated the student was at risk of failing the placement, IMS values were 6.65 and 9.90, with vastly different person measures of -.76 and - 6.84. In this case the IMS values were a more effective 'signpost' to indicate that this student's development of competency was at risk, rather than the absolute measure he/she received.

However, identifying the well fitting student with a Rasch score of 10.35 as at risk of failing is concerning as this student has passed the threshold score of 10.11 which indicates he/she would normally be considered to be competent. The original measurements of the ratings made on the VAS yielded ratings of 99 (5 items) or 100 (6 items). This may be due to a data entry error or CEs who provided the ratings considered that all 11 items should be rated at 100 (category 7). This does highlight that the Rasch model, which considers the threshold of competence being the point at which you have a 50/50 probability of scoring a 6 or a 7, may be problematic when considering the phenomenon of competence which assumes that all students must be rated at 7 to be considered competent. This will be discussed later.

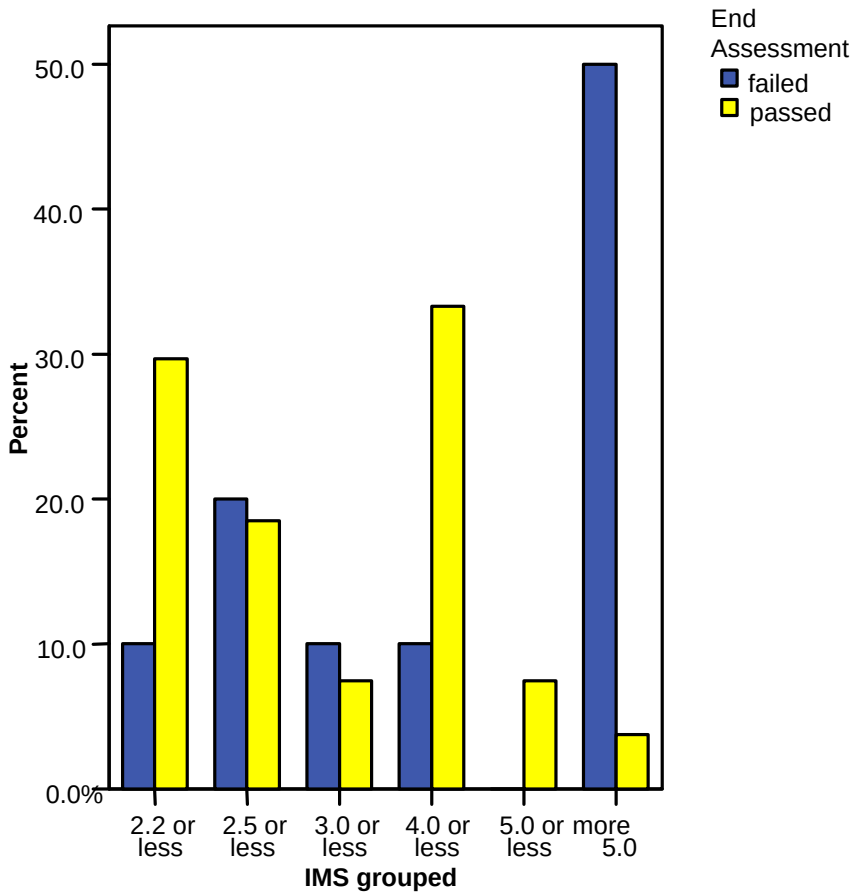
Table 34. Rasch Scores, Related Scale Category and IMS Values for Marginal Students for End Placement Ratings

Rasch Score	Scale Category (1 is lowest, 7 highest, +/- SE)	IMS
-11.57	1	1.05
10.35	7	1.17
-6.12	2	2.17
-.76	3	2.32
-7.74	2	2.37
-9.86	2	3.00
-2.46	3	3.42
-4.88	2	5.17
5.04	5 or 6	5.89
-7.74	2	5.95
-.76	4	6.65
-6.84	2	9.90

Determining Maximum Acceptable IMS Values

Data from marginal students (along with accepted Rasch analysis practice) suggests that an IMS value of above 2.0 may be an appropriate point at which to investigate whether a students' IMS indicates that their performance is highly variable, and therefore the person measure may not accurately reflect their ability level, or their performance may be marginal. Thirty seven students had an IMS value above 2.0 ranging up to 9.90 with person measures ranging from -9.86 to 9.12. The 27 students who were not identified as marginal had IMS values ranging from 2.10 to 5.23 with 8 falling between 2.0 and 2.2 as opposed to only one marginal student having an IMS below 2.2 (Fig 18). The peak of 9 non marginal students scoring between 3.1 and 4.0 is unexpected.

Figure 18. Pass/Fail rates of students with IMS values above 2.0



The rating strings generating IMS of 2.0 and above for non marginal students are provided in Table 37. The data was qualitatively investigated to see if the high IMS values could be accounted for according to three scenarios. This included whether students were performing higher than expected, students were underperforming and performances should have been examined to see if it was marginal, or if students had unexpected strengths/weaknesses. A fourth factor, idiosyncratic use of the rating scale by CEs could also be present but can not be teased out for this research design.

The category of performance in which students' Rasch score will place them can be identified by comparing their score (plus or minus their error score) with the thresholds on the VAS for each of the seven categories (as defined by Section 7.3. on calibrating the scale). These thresholds define the category ranges into which various scores fall and represent the groupings of scores that are so similar to each other that the instrument cannot separate these scores with sufficient degree of precision. Thus the scores that lie in the category or grouping defined by the step calibration fall into the same 'zone of competency'. The zones of competency were compared to hours of experience to determine whether students' performances (as represented by their scores) were placing them into a zone of competency

that could be expected given their level of experience. It must be noted that, while increasing hours of experience are strongly correlated with increasing Rasch scores, this is by no means a one to one correspondence as will be described later in Section 7.5.4. However, in the absence of any other yardstick by which to qualitatively evaluate the rating strings, Table 36 identifies what category levels are represented by the scores of those students who had IMS values of below 2.0 (and thus can be assumed to have accurate scores) grouped according to the hours of experience.

Table 35. Zones of Competency Represented by Rasch Scores with IMS Values Below 2.0 for Students Grouped According to Their Hours of Experience

Zones of Competency	Hours Category		
	0 to 80 hrs	81 to 180 hrs	180 hrs +
1	14	0	0
2	26	8	0
3	32	12	1
4	23	8	1
5	5	12	5
6	0	8	52
7	1	9	40
Total	101	57	99

Data in this table suggests that students with 0 to 80 hours of experience usually fall into zones 1 to 4 (94%) and students with more than 180 hours of experience end to fall into zones 6 or 7 (93%). However, as would be expected by the degree of variability identified previously (Section 7.5.3.3) regarding students with 81 to 179 hours of experience, their performances seem to be evenly spread across 6 of the 7 zones of performance. For the purposes of comparison it was decided to designate expected zones as those that could be reasonably expected for students of various levels of experience e.g. 1 to 3 for students with 80 or less hours of experience, as it seems unreasonable to expect students with such limited experience to reach zone of competence 4, even though some students clearly manage to perform at this level. The final categorisation used as a point of comparison for the rating strings as a qualitative strategy to highlight the range of variability that the Rasch analysis will identify as affecting the reliability of the Rasch score calculation is represented in Table 37.

Eight students had a Rasch score suggesting that they may have had performances above that expected for their hours of experience as indicated by a number of ratings higher than

would be anticipated for their degree of experience. Three students (6, 13, 24) had a Rasch score suggesting that the students were underperforming given the their hours of experience. These students' ratings were generally lower than would be expected for their degree of experience and suggest that the IMS value may have been indicating that closer attention was warranted regarding their overall performance to determine if in fact it fell into the marginal range. The other students with underfitting strings were characterised by either an unexpectedly low rating (2, 3, 9, 14, 26), very variable ratings across a wide range of competencies (7, 8, 16, 20, 25, 27), or an unexpectedly high rating (15, 21, 22).

These observations regarding the data are speculative in the absence of the opportunity to communicate with the raters or university programs involved. However, qualitatively there is evidence in the data to suggest that the three scenarios proposed above may be operating i.e. students who have specific strengths on one or two competencies, students who are overall performing highly, or students whose performance may in fact have been marginal but not flagged as such by the CE. The fourth scenario, of idiosyncratic rating patterns on the part of CEs continues to be a possibility as mentioned previously. It is also important to remember that some items are easier to gain higher ratings on than others (Table 33) and so will create variability in the raw data that is adjusted for by the Rasch model when calculating scores and IMS values.

Table 36. Rating Strings with IMS >2.0 for Non Marginal Students (n= 27)

Student.	Rating String	Rasch Score	IMS value	Zone of Competency by Rasch score	Est. Zone of competency by hrs.
1	45556666666	5.38	2.00	5 or 6	6 to 7
2	76765676676	9.12	2.04	6	6 to 7
3	44553235534	-0.17	2.06	4	3 to 4
4	22422223224	-4.88	2.10	3 or 4	1 to 3+
5	56565567766	6.57	2.10	6	3 to 5+
6	65645555465	3.75	2.14	5	6 to 7-
7	422421333*3	-4.07	2.16	2 or 3	1 to 3
8	34344334312	-2.74	2.19	3	3 to 4
9	66675576767	8.67	2.22	6	6 to 7
10	67766556665	7.06	2.26	6	3 to 4+
11	55555555525	2.51	2.30	5	3 to 4+
12	4545**44665	2.44	2.33	5	1 to 3+
13	55675556635	4.39	2.39	5	6 to 7-
14	322222321*2	-6.83	2.66	2	1 to 3
15	34252*24222	-3.81	2.86	3	1 to 3
16	55653245435	1.03	3.05	4	3 to 4
17	76676666646	7.61	3.05	6	3 to 4+
18	4442**354*5	-0.63	3.11	4	Unknown
19	**4632244*4	-1.19	3.12	3 or 4	1 to 3+
20	44435434413	-1.33	3.25	3 or 4	3 to 4
21	35453222322	-3.03	3.36	3	1 to 3
22	42424322222	-4.22	3.38	2 or 3	1 to 3
23	55775556366	5.04	3.73	5 or 6	3 to 4+
24	22522123232	-5.25	3.75	2	3 to 4-
25	224534241*2	-3.42	4.39	3	1 to 3
26	56675455777	6.57	4.64	6	6 to 7
27	27343323332	-2.46	5.23	3	3 to 4

Note. Zone of Competency by Rasch score represents the zone students' person measures fall into, plus or minus their personal error. Zone of Competency by hrs. is the zone students would be predicted to fall into given their hours of experience. The symbol '-'denotes predicted zone of performance by score lower than would be expected for hours of experience, '+' denotes higher performance. The symbol * denotes missing data.

Determining Minimum Acceptable IMS Values

As already noted, the data tends towards an overfitting pattern of ratings where strings of ratings given to students tend to more similar than proposed by the Rasch model. Bond and Fox (2001) recommend that, when determining minimum 'fitting' IMS values for items, values as low as .4 may be acceptable for judging tasks where agreement is encouraged, but do not comment regarding the setting of minimum IMS values for persons. As Curtis and Boman (2004) identify, the literature has focussed on fit statistics for items but is less informative on the topic of interpreting fit statistics for persons.

If an IMS value of below .4 is taken as a starting point, 43 persons in the data are represented. Low IMS values are not correlated with Rasch scores that range from -7.8 to 8.16 or with the level of CE's self rated experience. The data does suggest however that students in the lowest hours group for experience are disproportionately represented compared to the other two student experience groups. Twenty-four of the 41 students for which this information exists were in the lowest groups of hours with 8 students being in the middle group and 8 in the more experienced group.

This partially supports the previous suggestion that the lower group may be more consistent in their performance but it would have been expected to find more students from the senior group represented. This may be due to the hours range chosen for this grouping or because more CEs rating more experienced students have more categories below the students' average level of performance from which to select. The development of competence as evidenced by under and overfitting scores may form an interesting line of enquiry in the future.

There are two likely sources of erroneously overfitting data. First, the design and layout of the assessment tool encourages raters to be consistent in their rating patterns. Second, some raters may tend to rate their students at a similar point on the VAS across all competencies. This proposition appears to be supported by the fact that some CEs appeared to be overrepresented in the low IMS group. One CE contributed 9 sets of ratings (from a total of 20 submitted) that had IMS values below .4, another submitted 8 sets (from 15) and two submitted 4 sets each (from 10 submitted). All 4 of these CEs were employed by The University of Sydney supervising students in the lowest groups of hours so there may be some placement specific issues influencing the students' performance. It will not be possible to disentangle this issue without further research.

It was decided to not set a minimum IMS value and thus rely on the maximum IMS value of below 2.0 as sufficient for indicating that the person measure is accurate. This was decided due to the two possible factors described above creating low IMS values and the fact that Curtis and Boman's (2004) modelling of the effect of IMS values on the calibration of assessment scales suggested that low IMS values, as opposed to high IMS values, have little effect on the accuracy of the measurement. However, this is an issue that is worthy of further research and any future use of the assessment tool should keep in mind that a low IMS value (below .4) may indicate that an inaccurate person measure has been made.

7.5.3.4. Summary

Recommendations regarding suitable IMS ranges for this particular assessment tool have to be made on the basis of limited information in the Rasch literature on suitable IMS ranges for persons and the patterns associated with high and low IMS scores and the trends observable in the data. A conservative IMS range of 0 to 2.0 is suggested as the range within which person measures can be considered to be valid. An upper limit of 2.2 could also be justified from the data but a more conservative value of 2.0 is likely to assist in preventing marginal students, whose performances should be reviewed, from being overlooked. As mentioned, IMS values below .4 need to be investigated further to identify if a rater effect exists so it would be appropriate to monitor students receiving overfitting ratings and their raters with a view to ensuring that the person measures do reflect the students' levels of competence.

7.5.4. Person Measures and Relationships Within the Data

7.5.4.1. Process

Once person interval measures (logits) are determined by the Rasch analysis, a number of relationships within the data can be fruitfully examined between these measures and other characteristics or issues of interest. Given that persons with IMS values of 2.0 or above were identified in the above analysis as possibly having inaccurate person measures, these persons were removed from the data, leaving 265 cases available for analysis. However students with no IMS values due to having the minimum or maximum score for the assessment were retained for this analysis as they form an important part of the continuum of competence.

The interval person measures generated for each student and representing the ‘amount’ of competency each student possesses (plus or minus their standard error) enabled the relationship of person measures to a number of factors to be statistically examined using parametric statistics including:

1. Hours of student experience (cross sectional).
2. Student experience over time (longitudinal).
3. CE experience.
4. University attended.
5. Similarity of ratings from CEs assessing the same student.
6. Overall ratings by CEs.
7. Final placements.

7.5.4.2. Results

Hours of Student Experience

The relationship of person measures to hours of experience was investigated to confirm the hypothesis that student competence should increase with hours of experience and that scores on the assessment tool should reflect this. The hours of student experience (estimated and actual) also strongly correlated with the Rasch Score received (Pearson correlation = .823, p =.000). It is interesting to observe from the scatter plot of Rasch scores against hours of student experience (Fig. 19) that it is by no means guaranteed that students with more experience will receive an entry level score of 10.11 or above. Conversely, students with low levels of experience are represented in the entry-level score group of above 10.11. This suggests that competence is positively correlated with experience but not to the exclusion of other factors.

Figure 19. Scatter plot of student hours of experience against Rasch scores

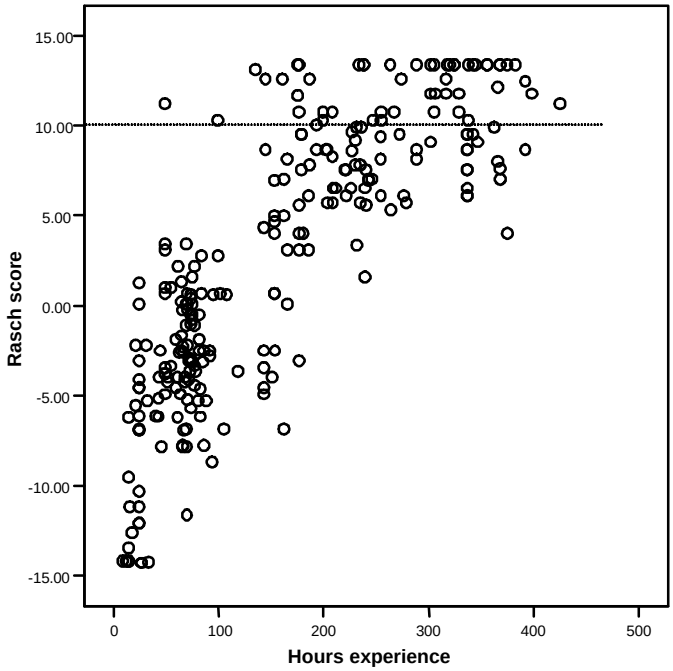
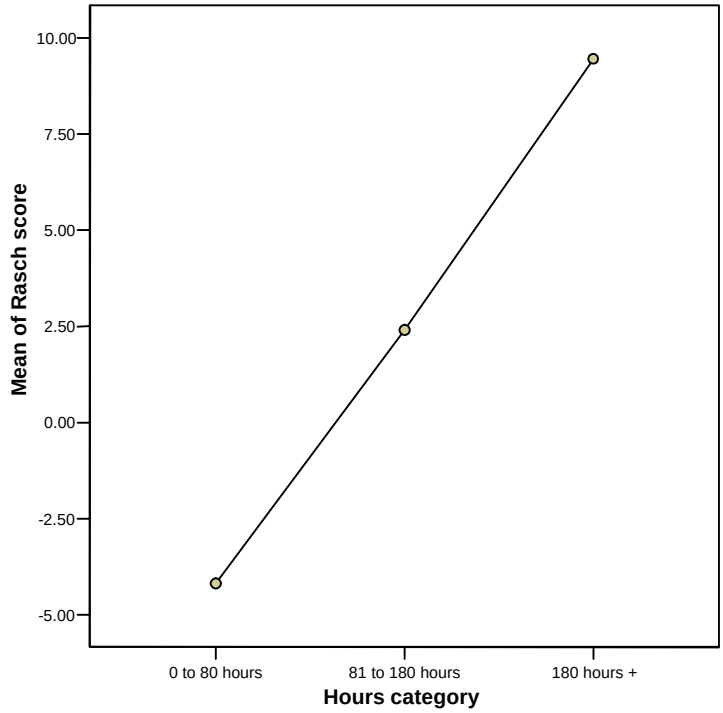


Table 38 clearly indicates that the mean score is highest for the group with the most hours and lowest for the group with the least hours, and that the amount of variance in the scores is highest for the middle group and lowest for the most experienced group. An ANOVA for these three groups identifies that these means are significantly different (p = .000) and the means plot (Fig. 20) illustrates the striking differences in person measures for each of these three groups. The relationship of competence to experience was further illustrated when the mean scores for students in their first placement (-8.23) were contrasted with the mean scores for students in their final placement (10.27). Not surprisingly, an independent sample t test for these two groups indicated that they were significantly different (p = .000).

Table 37. Means for Student Scores Grouped by Hours of Experience

Hours Group	N	Mean	St. Dev.
0 to 80 hours	102	-4.1829	4.77621
81 to 180 hours	58	2.4029	6.28119
180 hours +	98	9.4519	2.90275
Total	258	2.4767	7.55328

Figure 20. Means plots for students grouped by hours of experience



Student experience over time

A number of students had data on their performance submitted more than once over the 10 months the assessment tool was trialled. Once person measures with high IMS scores were removed, 20 students had a reliable person measure for more than one placement. Six students had 3 scores representing consecutive placements (coded placement 1 , 2 and 3), 10 had 2 scores for consecutive placements (coded placement 1 and 2) and 4 had 2 scores representing 2 placements separated by a placement for which no data was recorded (coded placement 1 and 3).

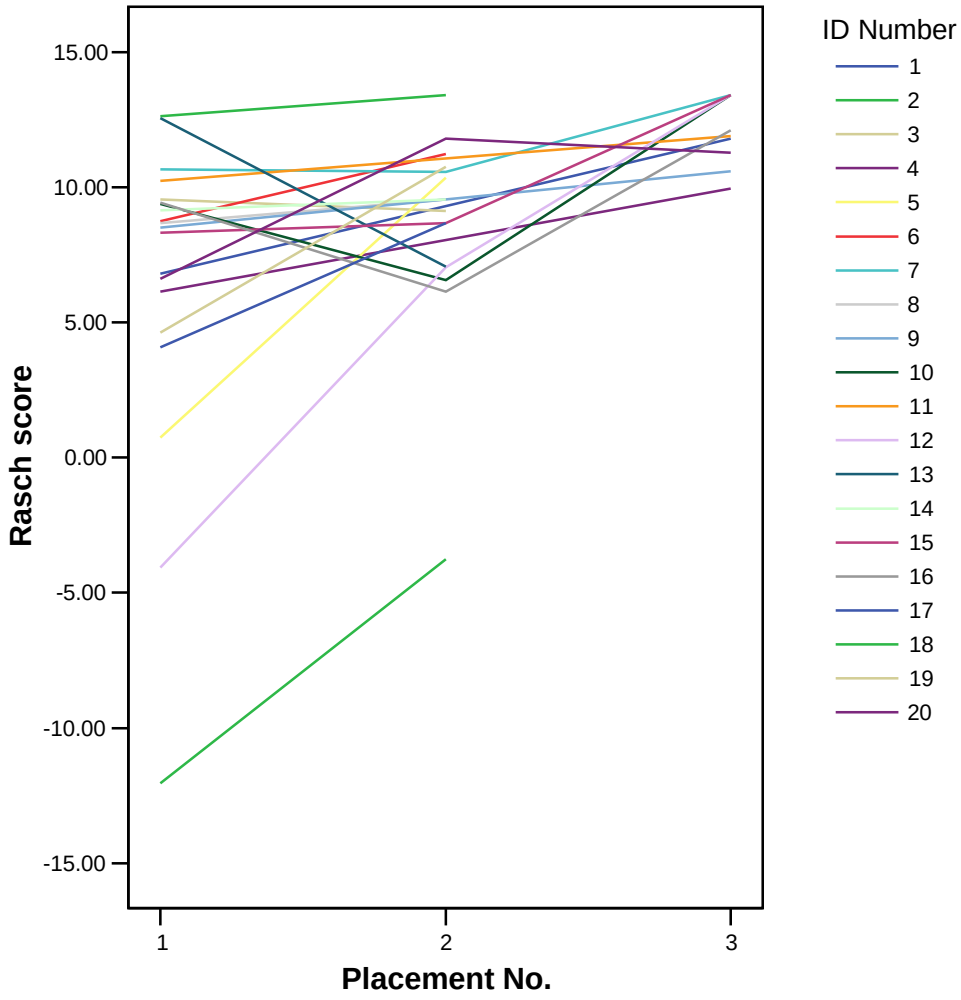
As can be seen from Fig. 21, the majority of students improved their competence score over time. Three students in this sub sample were already scoring above entry level competence (10.11 or above) or entered this range by the 2nd or 3rd assessment (14 students). Only 3 students had fluctuations in scores where subsequent scores were lower than the first person measure and could not be attributed to overlapping SE ranges.

Two of these students had 3 sets of scores (ID number 10 and 16 in Fig. 21) where the second score was lower than the first and third score. The third student had 2 sets of scores for consecutive placements and the second set was lower (ID number 13 in Fig. 21). The students with 3 sets of scores recovered and scored even higher on their third score than their first,

student number 13 with 2 sets of scores had the same CE as for student 10 who had a second score that was lower. Student 16 had a paediatric placement followed by a specialist adult placement followed by a second (but different) paediatric placement. With such a small set of data no strong inferences can be drawn. However, it is possible that some students' overall competence may not be as apparent in some placements compared to others, or they may strike a 'harder' rater and therefore be given a lower score, or even that the caseload was particularly challenging and this was not taken into account by the rater.

However, 17 of the 20 students represented in this sample clearly evidenced a steady increase in competence with subsequent placements regardless of having different raters and being placed with different clients and/or service delivery models. An ANOVA on the three groups indicated that this difference was significant at the .05 level ($p = .011$). This suggests that for these 17 students there were generic components to competence that develop with experience and regardless of placement type.

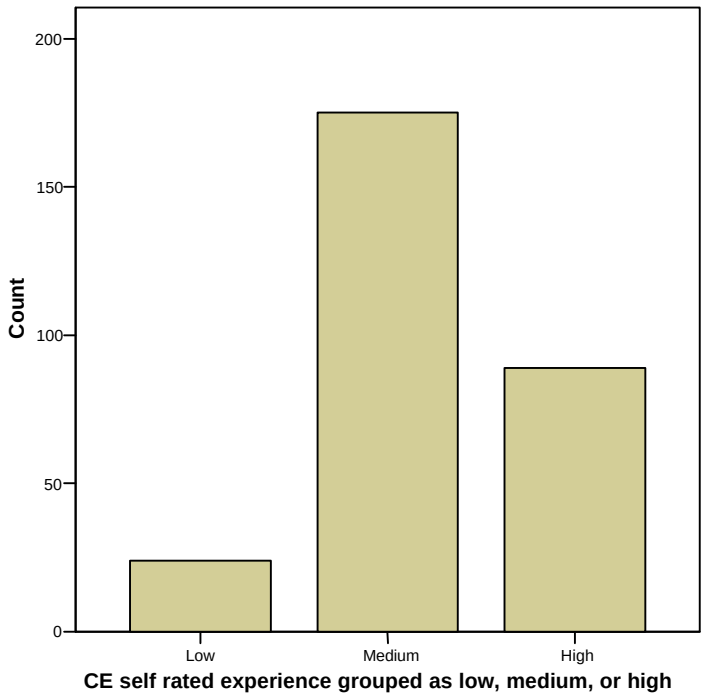
Figure 21. Rasch scores for students with more than one placement



Clinical educator experience

CE experience was investigated to identify whether experience affected the way in which CEs rated students on the assessment tool. As mentioned in Section 6.5.4. self rated clinical education experience seemed to be a better descriptor of experience than the number of students supervised and was used as a point of comparison in this analysis. It can be seen from this section that most CEs indicated their experience level as being 5 with very few in the lower two categories and relatively more in the higher categories. Given the generally small numbers in each category it was decided to collapse the categories into 3 groupings, representing: low experience (self ratings of 1 or 2); medium experience (self ratings of 3, 4 or 5); and high experience (self ratings of 6 or 7). Fig. 22 illustrates this grouping, which is somewhat subjective, but aimed to retain a grouping of CEs with little clinical education experience and separates out a grouping of CEs who identify themselves as being highly experienced. The distinction between high and low levels of experience is of particular interest as it has been suggested that the CEs' ability to make an accurate assessment of student performance is affected by the degree of previous experience they possess which forms a background against which comparative judgements are made e.g. Alexander (1996), Chapman (1998).

Figure 22. Self rated clinical educator experience grouped into three levels



The DIF function in RUMM (Andrich & Sheridan, 2003) was used to investigate whether there were differences in ratings between groups and which group(s) appeared to be rating in a different fashion. The DIF analysis indicated that ratings on 4 items were significantly different between the groups (Table 39) and the direction of these differences varied (Appendix 26 for DIF graphs). It became apparent however, that as the DIF graphs are plotted and compared over 5 class intervals, the already small 'low experience' group is subdivided further into even smaller groups for comparison meaning that less confidence can be held in the interpretation that ratings are affected only by CEs' experience. For example, for Item 9, the class interval (CI) groupings have very low CE numbers of 3, 2, 7, 7, and 5 from CI 1 to 5 respectively.

The differences in rating between the 3 experience groups appeared to resolve by the highest two CIs (Items 3, 7, 10) or the last CI (Item 9). The direction of variability differs for the four significant items; with inexperienced CEs rating students lower than other experience groups (Item 9, 10), higher than other groups (Item 3), or rating similarly to medium experienced CEs (Item 7) with more experienced CEs giving higher ratings for some CIs.

This preliminary investigation into the effect of CEs' experience on ratings suggests that differences may exist and may relate to expectations of performance on particular items rather than across the whole assessment. Further research with larger samples is required to identify if this is a real effect of CEs' experience or an artefact caused by low numbers of inexperienced CEs in the sample. If CEs' experience is found to have a significant impact upon rating patterns, this will usefully inform future revisions of the assessment tool and CE training and support. Regardless of this, the effect appeared to diminish as the students' competence level reached entry level thus suggesting that CEs were applying the entry level criteria in a similar way.

Table 38. Differential Item Functioning Analysis for 3 Levels of Clinical Educators' Self Rated Experience

Item Number	Competency Name	ANOVA: Degree of Self Rated experience (1 to 3)
1	GC Unit 1, Clinical Reasoning	0.177
2	GC Unit 2, Professional Communication	0.382
3	GC Unit 3, Lifelong Learning	0.047*
4	GC Unit 4, Professional Behaviour	0.077
5	CBOS Unit 1, Assessment	0.065
6	CBOS Unit 2, Analysis and Interpretation	0.805
7	CBOS Unit 3, Planning of Speech Pathology Intervention	0.040*
8	CBOS Unit 4, Speech Pathology Intervention	0.452
9	CBOS Unit 5, Planning, Maintaining, Delivering Speech Pathology Services	0.003*
10	CBOS Unit 6, Professional, Group and Community Education	0.040*
11	CBOS Unit 7, Professional Development	0.123

*Significant at a .05 level

University attended

The data was explored to identify whether students from different universities scored differently on the assessment tool and whether this requires consideration when applying the tool to different student groups. As more assessments were received from students attending The University of Sydney (180) than any other university, the assessments received from students from other universities were combined (84). The average person measures for The University of Sydney was 2.92 and students from the other universities was1.80. An independent samples t test for equality of means was conducted and found that these means were not significantly different (p = .258).

DIF analysis via RUMM was also undertaken to identify whether there were any differences at the item level that distinguished between the performances of students of similar ability levels according to the university program they attended. A number of different performances were identified (Table 40) suggesting that developmental progression towards competence on specific items may differ according to the university students attend. There are a number of factors that could be influential in this progression including how programs are

organised, shared expectations of student performance for particular ability levels, or a shared interpretation of how a particular item should be interpreted when rating. The DIF graphs (Appendix 27) indicated that these differences were generally not present for students of higher ability (CIs 4 & 5) suggesting that most resolve as the students approach competency. The factors influencing development of competency cannot be teased out in this current project but these results suggest that such analyses may provide useful information to universities about timing and sequencing of teaching and practicum experiences in relation to the development of competence.

Table 39. Differential Item Functioning for The University of Sydney vs. Other Universities Attended

Item Number	Competency Name	ANOVA: Uni. attended
1	GC Unit 1, Clinical Reasoning	0.497
2	GC Unit 2, Professional Communication	0.487
3	GC Unit 3, Lifelong Learning	0.067
4	GC Unit 4, Professional Behaviour	0.351
5	CBOS Unit 1, Assessment	0.000*
6	CBOS Unit 2, Analysis and Interpretation	0.566
7	CBOS Unit 3, Planning of Speech Pathology Intervention	0.032*
8	CBOS Unit 4, Speech Pathology Intervention	0.030*
9	CBOS Unit 5, Planning, Maintaining, Delivering Speech Pathology Services	0.000*
10	CBOS Unit 6, Professional, Group and Community Education	0.009*
11	CBOS Unit 7, Professional Development	0.214

*Significant at a .05 level

Similarity of ratings between CEs

An assessment tool that accurately assesses student competence would be expected to yield very similar ratings for the same student undertaking the same kind of work (client group and service delivery model) regardless of who is rating them (assuming similar opportunities for the CEs to observe and develop a judgement). Some variability would be expected owing to every client encounter and related tasks being unique and unrepeatable as well as the possibility that each CE will observe different numbers and occasions of speech

pathology practice by the student. However, if the assessment tool is successful in producing a repeatable and accurate person measure, ratings by CEs sharing a student placed in their workplace would be expected to be very similar.

As described in the methodology, joint CEs were actively sought by the researcher with the aim of receiving two sets of ratings from two different CEs teaching the same student in the same workplace and client group. This data proved particularly difficult to get, with far fewer students placed in these placements and joint CEs appeared to be more reluctant to participate in the research, citing workload issues. Once agreement to participate was gained there were further losses due to illness (2) and through one set of the paired data not being returned. In addition, the similarity of the placement experience for students was affected by CEs dividing the teaching task such that students experienced different types of service delivery models within the same organisation e.g. acute hospital care versus outpatient care or assessment clinic versus intervention services. Further to this one set of CEs indicated that, for 2 sets of data returned, the amount of time spent by each CE with each student observing/teaching was very dissimilar.

The data collection therefore resulted in 20 sets of data from students assessed by 2 CEs in similar but not identical workplaces experiences. If scores with high IMS values are removed, this is further reduced to 16 sets of data, with only 5 sets being from very similar placement experiences. These 16 sets of ratings yield a high intraclass correlation coefficient of .87 ($p = .000$). The intraclass correlation coefficient for the larger data set of 20 sets of ratings was also calculated, and was slightly lower at .83 ($p = .000$). Given the probable lack of similarity in placement experiences and amount of observations on which judgements were based, these correlation coefficients could be considered very satisfactory and suggest that CEs were highly likely to rate student performance very similarly. Thus the assessment tool enabled very similar person measures to be generated by each CE for their particular student.

A second and larger group of students had two assessments submitted by CEs working with them at the same time but in identifiably different placement settings e.g. university paediatric services versus school based programs. All clients were paediatric but differing in age and communication disabilities. An intraclass correlation coefficient was calculated on 33 students who had a score each from 2 different CEs with IMS values of below 2.0 and yielded a value of .82 ($p = .000$) which was not substantially different from that of students working within same workplaces. It is possible that an assessment community effect is operating, as all these students were assessed by CEs employed by The University of Sydney, and created very

similar ratings and therefore person measures for each student. On the other hand it may be that the elements of competency identified by the tool are observable to the same degree regardless of client group or service delivery model.

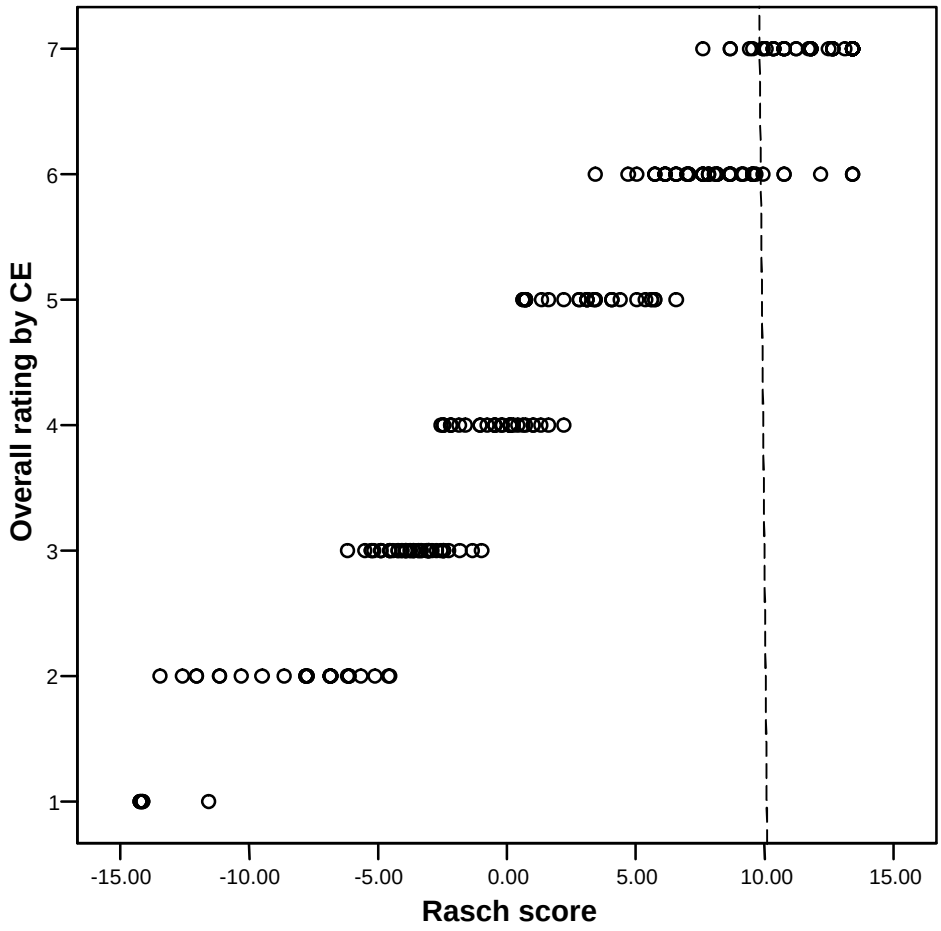
Given the constraints of the research design, the high intraclass correlation coefficients between raters suggested that the tool supports similar judgements by CEs about student competence. It is not possible to fully evaluate this aspect of the assessment tool without designing a judgement matrix and evaluating the effect of judgers upon the measurement reliability of the assessment tool with a two faceted Rasch analysis procedure as described by Linacre (1994; 1998a).

Relationship of global ratings to person measures

CEs were asked to give students a global or overall rating of competence on a VAS on the final page of the assessment tool. This rating was included with a view to evaluating the usefulness of a global rating as an indicator of the level of competence students had attained, particularly with reference to final students, and to examine the relationship of this overall assessment of competence with item ratings. Not all CEs provided this rating so the analysis is based on the 257 assessments that had a global rating and an accurate person measure (IMS below 2.0).

A scatter plot of the global rating category (indicated by converting the VAS measurement of the CE's rating into the rating category it represents) and the students' overall score was graphed. This indicated that as the students' overall rating category increased as did their overall Rasch score (Fig. 23). However, it can be clearly seen that there is overlap between the various categories of global ratings and associated person measures. This suggests that there is not a complete one-to-one correspondence between the degree of competency defined by summarising all ratings on the items and converting them into a Rasch score, and the overall global rating representing the CEs' judgement of competence.

Figure 23. Scatter plot of overall rating of competence against Rasch score



The overall category of rating given by CEs was a categorical rather than interval measurement so Spearman correlation coefficients were calculated to explore the relationship of ratings on the various items with the overall category of rating. Of interest was whether particular items may be more closely aligned with the global rating, which could suggest that they were more influential. As would be expected, all items are highly correlated with the global rating category given by CEs (Table 41). The item ratings that are most strongly correlated with the global rating of competence were: GC Unit 1 Clinical Reasoning; GC Unit 2 Professional Communication; and CBOS Unit 4 Speech Pathology Intervention. This group is closely followed by CBOS Unit 7 Professional Development, CBOS Unit 3 Planning of Speech Pathology Intervention, and CBOS Unit 5 Planning, Maintaining and Delivering Speech Pathology Services. The remaining 5 items did not appear to be as closely related to the overall rating given by CEs.

Table 40. Correlation of Global Rating With Item Ratings

Unit name	Correlation with overall rating
GC Unit 1 Clinical Reasoning	.945
GC Unit 2 Professional Communication	.945
CBOS Unit 4 Speech Pathology Intervention	.944
CBOS Unit 7 Professional Development	.942
CBOS Unit 3 Planning of Speech Pathology Intervention	.941
CBOS Unit 5 Planning, Maintaining and Delivering Speech Pathology Services	.940
GC Unit 4 Professional Role	.932
GC Unit 3 Lifelong Learning	.930
CBOS Unit 2 Analysis and Interpretation	.924
CBOS Unit 6 Professional, Group and Community Education	.925
CBOS Unit 1 Assessment	.916

Fig. 23 also indicated that there were a number of students (N = 8) of the 52 students who were rated in category 7 overall and had person measures below the cut off point (10.11) for a category or zone of competency level 7 rating on the assessment tool. Of these, 5 had scores with SE ranges that suggested that it was possible that their true score fell above 10.11. Three students had scores that suggested that they were outside the entry-level range even though their CEs had rated them overall in the 7th zone of competency and two of these had the same CE. These results could be due to idiosyncratic use of the VAS, such as end aversion bias. Useful information could have been gained by interviewing the CEs immediately after the assessment event to identify what processes were resulting in the discrepancy between overall rating and the zone of competency represented by person measure, but the research design did not allow for this.

Final placement students

The scores and ratings for students who appeared to have completed their final placement, as determined on the basis of the hours accumulated at the end of their placement and the dates of their placement period, were examined. If these students were at entry level, it would be expected that they would all receive a global rating of 7 and a person measure of 10.11 or

above (placing them in the 7th zone or entry level competency) unless identified as at risk of failing.

Table 42 describes all 37 students identified as completing their final placement and about to enter the workforce. Those students whose CEs indicated that their score may have been borderline for the cut off point of 'entry level competence' are indicated by the shaded boxes. Students whose score fell above 10.11 are in the boxes below the double line. Two students were identified as failing and a further four students had Rasch scores at or above entry level from a second and concurrent CE suggesting that a rater effect may have been present in these cases. However, this leaves 8 assessments from 7 students that placed them below entry level but passing, despite this being their final placement before graduation.

Table 41. Rasch Scores, IMS Values and Overall VAS Ratings for Students Completing Final Placements

ID number	Rasch Score	IMS Value	Overall VAS Rating (where provided)
31*	-4.88	5.17	43
75**	4.07	.58	N/A
2*	5.04	5.89	77
193	6.14	1.96	92
300	6.14	1.06	89
83	6.57	.64	91
273	7.04	.31	N/A
84	7.61	.85	92
175**	7.61	.09	91
139	7.63	.10	N/A
169**	8.06	.47	96
115	8.67	.86	100
272	8.67	.63	98
97**	8.67	.78	99
247	9.12	1.87	99
38	9.54	.92	93
176	9.54	.58	100
239	9.54	.63	96
68	9.95	1.26	98
159	10.35	.85	100

316	10.76	.88	100
72	11.23	.66	100
8	11.80	1.2	100
317	11.80	.81	100
170	12.18	1.08	97
98	12.49	.73	100
257	13.41	Max score	100
292	13.41	Max score	100
17	13.41	Max score	100
18	13.41	Max score	101
74	13.41	Max score	100
105	13.41	Max score	100
165	13.41	Max score	101
202	13.41	Max score	101
237	13.41	Max score	100
238	13.41	Max score	101
16	13.41	Max score	101

Note. Symbol * indicates CE identified performance as marginal, Symbol ** indicates that the joint CE rating (where available) yielded a passing Rasch score. Shaded area denotes zone where Rasch score plus or minus person error may place students' score at 10.11 or above.

One of these students, who had 2 assessments represented in this group (83 and 84), was placed in her first and only adult neurological placement (A. Russell, personal communication, May 2004) and this was likely to be the reason for her not reaching entry level on this placement, despite it being the final placement prior to graduation. However, a second student was in the same position and shared one of the CEs of student 83/84 who also gave her a score below the cut off (175). However, the second CE gave this student a score within the cut off zone (176) so it would appear to be possible to reach competency with this degree of experience with the client group. Without further qualitative investigation it is not possible to determine what factors influenced the CE to deem this student as not being 'at risk of failing this placement' given the expectation that students need to reach entry level on their final placements. However, it does suggest that the score yielded by the assessment may provide extra information to assist in this decision making process.

There were a further 6 students with person measures below entry level (193, 300, 273, 139, 115 and 272) and their raw VAS ratings (0 to 100 units, 101 representing above entry

level) were investigated to determine whether end aversion bias may have been operating (Table 43). CEs appear to be using the end of the scale or the above entry level box quite freely. The raw ratings suggest that these 6 students may be a group whose performance needed to be carefully evaluated before being deemed entry level and permitted to graduate. Of the 4 out of the 6 students who had an overall rating from the CE, 3 were given an overall rating of below entry level (99 or below), however none of the CEs identified their student as 'at risk of failing this placement'. Again, it may be that other valid factors influenced the CEs' decision to not identify the student as being at risk including the possibility that this was not seen as the placement CE's responsibility to determine. The person measure generated by the assessment would have provided a useful marker for the coordinator of the practicum program that this decision may require further evaluation before a final placement student is deemed as being at entry level. The person measure would suggest that the student has not reached entry level or the CE requires training regarding expectations for entry level competence.

Table 42. VAS Ratings for Students Deemed Overall Competent by Their Clinical Educator on Their Final Placement With a Rasch Score of <10.11

	Student ID Number					
Competency	193	300	273	139	115	272
GC 1	90	85	94	96	96	99
GC 2	87	88	*	*	97	99
GC 3	90	89	91	91	98	100
GC 4	101	88	93	93	97	95
CBOS 1	95	90	91	91	98	90
CBOS 2	83	90	86	92	97	92
CBOS 3	82	92	91	91	99	99
CBOS 4	101	97	89	92	99	100
CBOS 5	78	86	90	90	100	92
CBOS 6	78	97	90	93	99	92
CBOS 7	92	86	90	93	100	99
Overall Rating	92	89	*	*	100	98

Note. Symbol * denotes missing rating.

7.5.5. Summary

Investigating the person measures generated by the Rasch analysis confirmed that the assessment tool is very robust. The assessment tool clearly measured change in competence

according to hours of experience and also over time for individual students. Intra class correlations of the ratings provided by two different rater groups for their shared students indicated that the assessment tool yielded very similar person measures for students regardless of who provided the ratings.

In addition, the assessment tool was sensitive to a number of important features of student performance on the rating items. IMS values of above 2.00, using the anchored thresholds determined during calibration, proved to be an excellent marker of students whose competence was marginal due to very variable performances or who showed specific strengths and weaknesses. This also confirmed the subjective assessments made in the literature that marginal students are frequently inconsistent in their performances (Robertson et al., 1997). This IMS value has the potential to be very useful as an indicator that the students' person measure may accurately reflect their actual level of competence and that their performance requires careful review and consultation with the CE providing the ratings. The person measures also clearly provide a benchmark for entry level performance against which the overall judgement of a student being at entry level can be compared and evaluated.

Other interesting aspects of analysis of the person measures included findings that suggested that CEs' global judgement of competence may be more influenced by some competencies than others. Competency also appears to have sufficient generic components that may enable students to perform similarly across different client groups and service delivery models at the same point of time. It was clear that competence does develop with experience but experience is not the sole influence on this development. Finally, the tool may also provide information that reflects the way in which different programs influence the development of their students' competence

The final analysis undertaken to evaluate the validity of the assessment tool was examination of feedback data from students and CEs who used the tool and is described in the next chapter.

CHAPTER EIGHT

8. USER EVALUATION OF THE ASSESSMENT TOOL

Feedback from the students and educators using the assessment tool was sought to provide further information on the validity of the tool. This chapter describes the feedback provided by users regarding their experience of the research assessment tool and their perception of its face validity. This information was used to identify factors that might affect its reliable and valid use, provided useful information for evaluation of the validity of the tool, and identified issues to consider in future tool revisions.

8.1. Method

The questionnaire was designed according to principles identified that maximise effectiveness of consumer evaluation surveys (McAllister & Brown, 1999). Respondents were asked to rate their agreement to statements using a 7 point scale (1 = Strongly Disagree and 7 = Strongly Agree) so the strength of the agreement could be indicated and to allow for a neutral opinion of '4'. Both positive and negative statements were used to avoid a 'response set' occurring. Attention was directed to ensuring that statements were worded so that the common meaning was clear and that one issue only was addressed by each statement. Open-ended questions were included to solicit information that may otherwise not have been provided and points of prime interest were addressed through sets of statements rather than single measures.

Students and their CEs participating in trialling the prototype assessment tool were invited to complete feedback sheets and return completed forms via a reply paid address. Two versions were provided, each containing similar statements worded according to the CEs' or students' perspective and addressing several aspects of tool design, content, assessment process, and research procedure (Appendix 28). The CE version had 26 items and the student version had an extra item under the group of statements related to the rating scale (see Section 8.3.1. below). Demographic information was also collected as well as information on other aspects of the assessment tool.

The feedback questionnaire covered the following topics:

1. Design and utility of the visual analogue rating scale used to record student competency.

2. Clarity and validity of the behavioural descriptors developed to guide the use of the rating scale.
3. Structuring the assessment so that a more detailed assessment was required at mid placement than at the end of the placement.
4. Usefulness of the resources provided to support use of the research assessment tool.
5. Relevance and usefulness of the generic competencies developed during the research project as assessment items.
6. Usefulness of the prototype tool when working with marginal students.
7. Information regarding the use of the hard copy versus online version of the assessment.
8. Clarity of the research procedure and availability of support.
9. Validity aspects of the tool such as how well the research tool supported the CEs'/students' judgement regarding their competency and the assessment and teaching/learning process.
10. Overall satisfaction with the research assessment tool.

8.1.2. Demographics

A good response rate from CEs was achieved with 68 of the 107 (64%) participating CEs returning a feedback form and a satisfactory return rate of 88 from the 219 students (40%) was also achieved. Not all feedback sheets were complete. Feedback questionnaires were returned by CEs and students from a variety of university programs, student year levels, and placement types as can be seen by Tables 44, 45, and 46. Feedback questionnaires were matched with demographic data from the assessment tool to identify CE and student experience where this information was available. Figs 24 and 25 illustrate that a range of experience was represented in the feedback sample.

Table 43. University Program Represented by Clinical Educators and Students Providing Feedback

University	CEs supervising students from this University(s)	University attended by students providing feedback
The University of Sydney	27	49
The University of Newcastle	4	4
Charles Sturt University	4	5
Macquarie University	4	6
La Trobe University	3	2
Flinders University	17	17
The University of Queensland	9	4
Not stated	0	1

Table 44. Student Year Levels Represented by Clinical Educators and Students Providing Feedback

Year Levels	Clinical educators by yr. level of student in placement	Students providing feedback
Yr 2	N/A	2
Yr 3	16	32
Yr 4	39	48
Masters Yr 1	2	5
Masters Yr 2	2	1
Yrs 2 & 4	2	N/A
Yrs 2, 3 & 4	5	N/A
Yrs 3 & 4	1	N/A
Not stated	1	0

Table 45. Numbers of Weeks and Days per Week of Work Placement Provided by the Clinical Educators and Undertaken by Students Providing Feedback

	No. of days at placement							
	1 day		2 days		3 days		Block (4 or 5 days)	
No. of Wks. at placemt.	CE	Student	CE	Student	CE	Student	CE	Student
4							1	
5							3	2
6			1				19	21
7	1	1	1					
8			1		4	4	2	5
9		1	2	3			1	
10		3	5	3			4	5
12	7	10	1	4				
13	4	14		3		3	1	
14	1							
17					1			
18	1							

Figure 24. Degree of self rated expertise for those clinical educators who provided feedback (N=59)

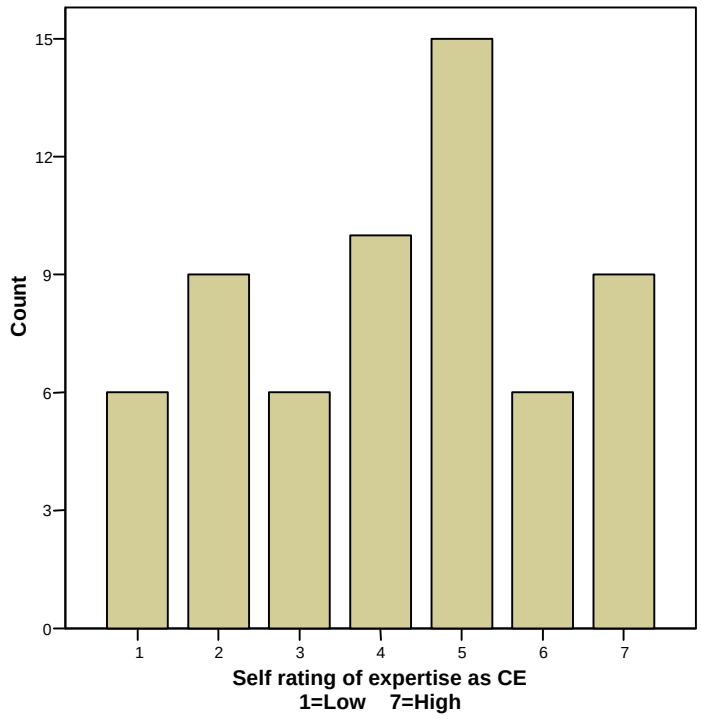
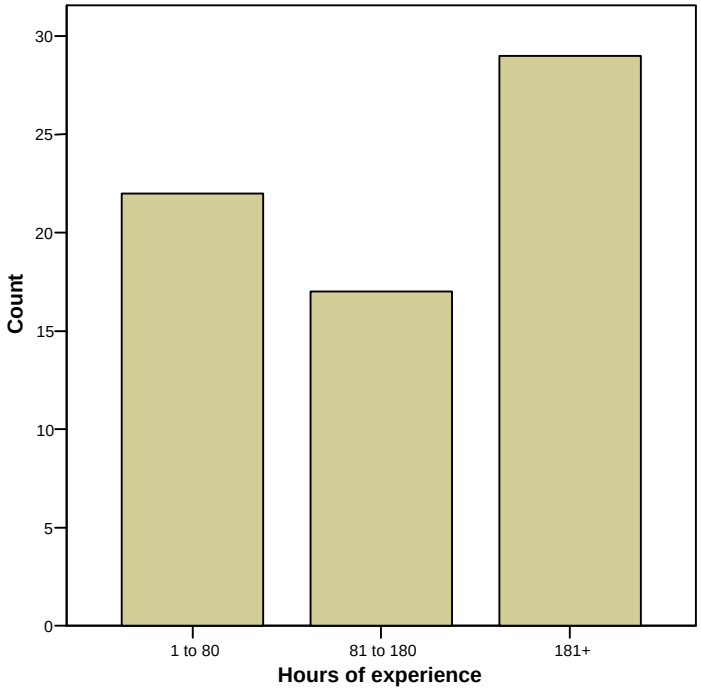


Figure 25. Hours of experience of those students who provided feedback (N=68)



8.1.3. Time Taken to Complete Assessment

CEs were asked to indicate the time it took to complete the assessment tool at mid and end placement as an indication of how practical the task was to carry out. The time taken to complete the assessment tool for students at mid placement varied considerably (15 to 180 minutes) with the mean amount of time taken to complete the mid placement assessment being 62 minutes and the mode being 60 minutes. The briefer end placement assessment averaged 32 minutes (with a range of 5 to 90 minutes) but the mode was 15 minutes.

These times need to be interpreted with some caution as not all CEs provided this information, comments indicated that some times included completing the usual university assessment as well as the research tool, and assessments of students identified as marginal took considerably longer. In addition, some assessments were rated jointly with the student, this presumably would have created discussion and taken longer than doing the assessment without the student present. Three CEs commented in their feedback that the assessment tool was time consuming but got faster with familiarity and two indicated that they found it faster if they did not fill in the comments sections provided. In addition the University of Sydney CEs were required to use the online version as well as print off a hard copy for each student's record which was very time consuming as the online version was not designed for this purpose. This issue could be easily managed with changes to future design specifications.

No-one specifically complained that the time the assessment tool took was onerous. Comments were also made regarding the usefulness of the more detailed mid placement assessment as a teaching tool and approving of the brief end placement assessment (this will be elaborated on further). A CE in the field commented that *"[the online version] seriously sped up the process I'm sure."*

8.2. Analysis

The 26 items of the CE feedback questionnaire that were rated by the majority of CEs were subjected to a Rasch analysis using Bigsteps (Linacre & Wright, 1998) and the rating scale analysis procedure outlined in Section 7.2, Chapter Seven in this thesis. The process aimed to identify whether the questionnaire met the assumptions of the Rasch model and thus functioned as a measurement tool that quantified the relative amounts of satisfaction CEs possessed regarding the research assessment format. If the questionnaire functioned in this manner for CEs, a similar process would have been used to evaluate student ratings. It became apparent that the rating scale could not be resolved satisfactorily into a single set of well functioning categories for all items on the scale in a manner that provided reliability statistics above .80 for this particular sample.

The rating of 4 consistently had a disordered threshold (step calibration) and the probability curve for this category was never modal, indicating that at no point on the variable was it the most likely category to be observed (Linacre, 2002). Thus it appeared to be functioning as a 'neutral' or 'no opinion' selection rather than as a definable section of a continuous variable of satisfaction with assessment tool.

The categorisation that most closely approached the Rasch model requirements was collapsing ratings 1, 2, 3 into one category, omitting ratings of 4, collapsing ratings of 5 and 6 into one category and designating rating 7 as the final and 3rd category. This yielded a person reliability of .79 and item reliability of .87. However, 3 negatively rated items were misfitting (OMS above 1.4) and the observed percentage coherence statistics from Category 1 and 3 were poor (22% and 30% respectively). This coherence statistic suggests that it cannot be assumed that the ratings imply a measure as there were more ratings in these categories than would be predicted to be there if the feedback questionnaire was a precise measurement tool or ruler for measuring satisfaction (Linacre, 2002). There would appear to be a large degree of variability within the data indicating that respondents cannot be assumed to respond

predictably according to their suggested degree of overall satisfaction (person measure) but were responding differentially according to the item they were rating.

A number of other analyses were conducted, including deleting misfitting items and continuing with the 3 category solution above and reducing the data to a dichotomous categorisation of satisfied/not satisfied for all items. The dichotomous categorisation was trialled for all 26 items but was found to produce poor person reliabilities (.40) and item reliabilities less than .80 (.74). Coherence statistics continued to be poor for the negative category (observed 36%).

The majority of misfitting items for the 3 category solution were those that were negatively stated and thus requiring respondents to reverse the rating to indicate satisfaction. Removal of the first three with high OMS values resulted in more misfitting negatively worded items being identified in the reanalysis. Nine items had to be removed before all items had OMS values below 1.4. Even with these items removed the coherence statistics were poor for the remaining items and, while the person reliability remained the same, item reliability had dropped to .75 and a greater than 5.0 logit gap existed between the thresholds (5.86 logits) indicating poor measurement properties.

The negatively worded items were analysed as a separate group and found to have poor person reliabilities of .41. This indicates that the satisfaction levels of respondents to the questionnaire were not well targeted by the items and the ordering of person satisfaction was highly unlikely to be maintained if a set of items measuring satisfaction in a similar fashion were administered (Bond & Fox, 2001). This may have occurred because some respondents did not alter their response set and continued to select higher rating numbers to indicate satisfaction or because responding to negatively worded items is a different task resulting in different patterns of ratings (Mosenkis, 1997).

Overall the Rasch analysis suggested that, while the feedback questionnaire provides a vehicle for CEs to express positive, negative, or neutral opinions regarding specific aspects the assessment tool, it did not function as a reliable ruler specifying quantifiable amounts of satisfaction amongst CEs. This finding simply implies that the items represent more than one variable presumably related to satisfaction and that CEs' responses to items indicated that respondents based their satisfaction with the tool on different aspects of the tool design and use. Given all items were considered to identify important aspects of tool design and usage, it was not judged to be appropriate to eliminate any items. The analysis did suggest that a rating of 7 represented a genuinely higher level of satisfaction than a rating of 5/6 but not to a degree

where strong confidence could be held in this interpretation. Thus a qualitative rather quantitative analysis of the data divided into negative/neutral/positive ratings was the most justifiable interpretation of the ratings given. Finally, it appeared that negatively worded items tended to attract more negative and strongly positive ratings than positively worded items, requiring that these categories must be interpreted with caution for these items.

8.3. Results and Discussion

Based on the Rasch analysis described above, ratings by CEs and students were collapsed to indicate agreement (rating a 5, 6 or7), neutral (4), or disagreement (1, 2, 3). Thirty three students and 51 CEs provided comments. The pattern of rating regarding the research assessment tool was positive overall. CEs and students agreed with positive statements and disagreed with negative statements (Tables 50, 52, 53, 54, 55 & 56 under subsequent sections). This gave useful insight into how well received the research tool was and areas that needed consideration.

A Mann-Whitney U test was calculated and indicated that CEs and students rated in a similar pattern except for three statements (Table 47). Students rated in a less positive pattern regarding the usefulness of the examples in the assessment resource manual of applying the behavioural descriptors to the competencies. It is likely that the research methodology used resulted in CEs having access to the resource materials and using them for the assessment but not necessarily the students. The research process may also have resulted in students giving less positive and more neutral responses regarding the usefulness of the assessment tool in goal setting. Students were significantly more positive than CEs about the usefulness of the greater detail at mid semester in relation to their learning. However, as CEs were answering this in relation to their teaching, this is probably due to the question being, in effect, entirely different for each group and so not directly comparable.

To investigate the impact of experience on perceptions of the tool, correlation coefficients were calculated between ratings and CE and students' experience as represented by self rated degree of experience in clinical education and hours of experience respectively (Tables 48 & 49). A number of significant correlations were found and will be discussed in the relevant sections below.

Table 46. Feedback Statements With Significantly Different Rating Patterns Between Students and Clinical Educators

Feedback Statement rated by clinical educator/*student*	Mann-Whitney U	Z score	Asymp. Sig. (2 tailed)
The examples in the Assessment Resource Manual (page 13) of how the behavioural descriptors might apply to the CBOS and Generic Competency Units were useful.	1942.50	-3.41	.001**
Having more detail in the form of ratings on elements at mid placement helped me with my teaching/*learning*	2333.00	-2.44	.015*
The research Assessment Tool did not help in goal setting with the student(s) /*did not help me goal set*.	1721.50	-1.99	.047*

** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

Table 47. Clinical Educator Ratings Significantly Correlated With Self Rated Degree of Experience

Statement rated by Clinical educator	Corr. with self-rated experience. (Spearman's)	Sig. (2-tailed)
The Rating Scale was difficult to use. N=61	-.272 (*)	.041
I felt confident about making my judgment based on the Rating Scale. N=61	.295(*)	.023
The examples in the Assessment Resource Manual (page 13) of how the behavioural descriptors might apply to the CBOS and Generic Competency Units were useful N=61	-.384(**)	.007
The student(s) should also be rated on elements at end placement. N=61	-.233	.076
The Generic Competencies reflect knowledge, skills, and attitudes of value to the profession. N=59	.448(**)	.000
The GCs were an unnecessary inclusion in the assessment. N=59	-.280(*)	.035
The GCs were a good description of the competencies that underpin competent practice of speech pathology. N=60	.373**)	.004

** Correlation is significant at the .01 level (2-tailed)
* Correlation is significant at the .05 level (2-tailed)

Table 48. Student Ratings Significantly Correlated With Hours of Experience in Speech Pathology Practice

Statement rated by Student	Corr. with hrs. of experience. (Spearman's)	Sig. (2-tailed)
The Rating Scale reflected my progress well. N=73	.348(**)	.004
I would rather use a Categorical Scale. N=72	-.271(*)	.0026
The Rating Scale was effective in showing my progress over time	.33 (**)	.006
The behavioural descriptors do not match my understanding of how competence develops. N=73	-.404(**)	.001
Students should also be rated on elements at end placement. N=72	-.212	.085
The research Assessment Tool did not effectively support my judgement of my competency. N=71	-.364(**)	.003
The research Assessment Tool is effective in identifying my strengths and weaknesses.	.348(**)	.004
The research Assessment Tool did not help me goal set. N=53	-.500(**)	.000
How would you rate your OVERALL satisfaction with the research Assessment Tool as an assessment of your competency in your clinical placement?	.276(*)	.029

** Correlation is significant at the .01 level (2-tailed)
* Correlation is significant at the .05 level (2-tailed)

8.3.1. Rating Scale

The research assessment tool used a visual analogue scale (VAS) for ratings rather than the categorical scale currently used by Australian universities. All respondents felt that the rating scale reflected students' progress well and generally felt confident about making a judgement on the scale (Table 50). In addition, 88.6% (N= 78) of students indicated that the rating scale reflected their progress well, with more experienced students being significantly more likely to agree with this statement. More experienced students were also significantly more likely to disagree that they would rather use a categorical scale and more likely to feel

the VAS was effective in showing their progress (Table 49). However, relatively fewer respondents were positive about its ease of use. Most respondents disagreed that they would rather use a categorical scale with quite a few expressing a neutral rating of '4'.

Table 49. Clinical Educators' and Students' Feedback Ratings in Response to Statements Regarding the Rating Scale.

Statement rated by clinical educator / *student*	Response category and median rating	CE (% resp.)	Student (% resp.)
The Rating Scale used reflected the student's / *my* progress well.	Agreed	86.6	88.6
	Neutral	9.0	5.7
	Disagreed	4.5	5.7
Clinical Educator N = 67 Student N = 88	Median rating	6	6
The Rating Scale was difficult to use.	Agreed	18.2	20.5
	Neutral	13.6	17
	Disagreed	68.2	62.5
Clinical Educator N = 66 Student N = 88	Median rating	2	3
I felt confident about making my judgement based on the Rating Scale.	Agreed	73.1	73.3
	Neutral	16.4	12.7
	Disagreed	10.4	14
Clinical Educator N = 67 Student N = 86	Median rating	5	5
I would rather use a Categorical Scale. (Such as the one used for this question).	Agreed	28.4	36.8
	Neutral	22.4	19.5
	Disagreed	49.3	43.7
Clinical Educator N = 67 Student N = 87	Median rating	4	4
The rating scale was effective in showing my progress over time.	Agreed	*Not rated*	85.2
	Neutral		9.2
	Disagreed		5.6
Student N = 87	Median rating		6

Eight students specifically commented regarding concerns with the VAS format of the rating scale, generally indicating lack of confidence about where to rate oneself on the line. This concern was probably reflected by the number of students who indicated that they preferred a categorical scale (36.8%, N=32). Two students also commented that it was hard to rate oneself with regard to entry-level when they were not really clear as to entry level expectations. Five CEs also indicated that they did not like the VAS format for similar

reasons e.g. *"I and most of my students found the visual analogue scale too ambiguous at times, and would prefer a categorical scale."* On the other hand, three CEs and 1 student indicated that they preferred the VAS format.

This feedback would suggest that most CEs and students were receptive to the VAS format but some lacked confidence in using it. This may be due to its unfamiliarity in this context and lack of training in applying it to the assessment of competence in speech pathology. This is further supported by the finding that more experienced CEs were significantly more likely to indicate that the Rating Scale was easy to use and to feel confident about making a judgement based on the rating scale (Table 48). This suggests that experience contributes to confidence and may be at least as important as the assessment format itself.

8.3.2. Behavioural Descriptors

Descriptions of behaviours indicating novice, intermediate, and entry-level performance and detailed examples of how to apply these descriptors to the competencies being rated were provided to support valid and reliable rating on the VAS. Responses indicated that these were easy to understand and matched respondents' understanding of how competence develops (Table 51). More experienced students were significantly more likely to indicate that the behavioural descriptors matched their understanding of the development of competence (Table 48). More CEs found the examples useful than did the students, although 21% (N=17) of students responded neutrally which may indicate that a number did not see the resources. Only 16% (N=14) of students and 22% (N=15) of CEs agreed that the descriptors were difficult to use when judging levels of competence. The three questions that related to resources for use of the behavioural descriptors under question 4 "Resources in the Assessment Resource Manual" were very positively endorsed by CEs and students (Table 52).

While the large majority of students rated the behavioural descriptors as 'easy to understand' there were four comments indicating that students found them or aspects of the assessment tool unclear or containing too much jargon. Overall students had a lower median rating for 3 out of 4 items regarding the behavioural descriptors, indicating that they may have found the behavioural descriptors harder to understand than CEs.

The feedback confirms that the behavioural descriptors were generally understandable and appropriate to the judgement of competency (important considerations for reliable and valid use of the scale) and that the resource material was useful. Again lack of familiarity and training on the new assessment tool may explain the users' lack of confidence in applying the

behavioural descriptors when assessing levels of competency. This may also explain why less experienced CEs were significantly more likely to have rated the resource materials and the usefulness of the examples of applying the behavioural descriptors more positively than more experienced CEs (Table 48).

Table 50. Clinical Educators' and Students' Feedback Ratings in Response to Statements Regarding the Behavioural Descriptors

Statement rated by Clinical Educator / *Student*	Response category and median rating	Clinical Educator (% responding)	Student (% responding)
The behavioural descriptors were easy to understand.	Agreed	88.1	76.1
	Neutral	4.5	10.5
	Disagreed	7.5	12.8
Clinical Educator N = 67 Student N = 86	Median rating	6	5
The behavioural descriptors were difficult to use to judge the student's/*my* level of competence.	Agreed	22.4	16.2
	Neutral	9.0	17.4
	Disagreed	68.6	66.4
Clinical Educator N = 67 Student N = 86	Median rating	3	2.5
The behavioural descriptors do not match my understanding of how competence develops.	Agreed	5.9	7.9
	Neutral	8.8	10.3
	Disagreed	85.3	81.7
Clinical Educator N = 68 Student N = 87	Median rating	2	2
The examples in the Assessment Resource Manual (page 13) of how the behavioural descriptors might apply to the CBOS and Generic Competency Units were useful	Agreed	91.2	71.1
	Neutral	7.4	20.5
	Disagreed	1.5	8.4
Clinical Educator N = 68 Student N = 83	Median rating	6	5

Table 51. Clinical Educators' and Students' Feedback Ratings Regarding Usefulness of Resources in the Assessment Manual Providing More Information on the Behavioural Descriptors

Statement rated by Clinical Educator / *Student*	Response category and median rating	Clinical Educator (% responding)	Student (% responding)
Resources: Behavioural Descriptors (assessment tool)	Agreed	83.9	83.2
	Neutral	9.7	11.2
	Disagreed	6.5	5.6
Clinical Educator N = 62 Student N = 71	Median rating	6	6
Resources: Behavioural Descriptors - Detailed version in Resources Manual	Agreed	84.2	76.8
	Neutral	8.8	13.9
	Disagreed	7.0	9.3
Clinical Educator N = 57 Student N = 43	Median rating	6	6
Resources: Examples of applying the Behavioural Descriptors to the CBOS and Generic Competencies	Agreed	89.7	76.2
	Neutral	10.3	16.7
	Disagreed	0	7.1
Clinical Educator N = 58 Student N = 42	Median rating	6	5

8.3.3. More Detail at Mid Placement and Summary Assessment at End Placement

The prototype tool was designed so that more detailed ratings were made at mid placement than at the end placement. Most current assessments have the same ratings at mid and end placement ratings, or more at the end placement ratings. Good educational practice involves providing more detail at formative assessment (mid placement) than at the summative assessment (end placement). Respondents agreed that having more detail at mid placement assisted with their teaching and learning and did not feel there was too much detail provided (Table 53).

The majority of the 58 CEs who responded to the questions "Did you refer to the Element Level (*as in the Mid Placement Assessment)* as well as the Unit Level when making the end of placement judgement of the student's competency?" indicated 'yes' (62%, N=36). Only 25% of the 80 (N=20) students who responded to this question indicated 'yes'. CEs who responded to the question regarding which elements they referred to at end assessment generally

indicated all of the elements were reviewed at least briefly (27 comments) to support their final judgement of the student or for teaching and learning purposes:

> *"I tended to refer to most of them, particularly so I could give my student feedback about specific areas, and comment on improvements from the mid placement."* (CE Feedback 43)

Nine students who responded to the general invitation for feedback at the end of the questionnaire commented positively regarding the amount of detail involved in the assessment, with two commenting specifically about mid placement assessment.

> *"It was good to have detailed feedback after mid placement assessment because it gave me the opportunity to find out how I was going and where my strengths and weaknesses were. I feel this assessment procedure has given me a lot more feedback than the [usual university assessment]."* (Student feedback 68)
> *"Very comprehensive and gave much direction and opportunity for feedback."* (Student feedback 13)

Opinion was much more divided about whether the same number of ratings should occur for each assessment. However, when correlated for CE and student experience, it became apparent that with more experience both CEs and students were significantly more likely to disagree that detailed ratings were required at the end of placement (Tables 48 & 49), lending weight to the decision to structure the tool in this way. A number of CEs comments indicated that they found it sufficient to review the list of elements under the unit heading in the research tool before making their ratings, with some suggesting that more detailed ratings should be an option if required e.g. if the student is marginal.

> *"All [elements were looked at]! It was useful to refer to, but I don't think it needs to be included in end [assessment as in the mid-placement assessment]."* (CE feedback 56)

A student also commented:

> *"…liked shorter end placement but should be option to rate all competencies if required…"* (Student feedback 15)

Table 52. Clinical Educators' and Students' Feedback Ratings in Response to Statements Regarding the Level of Detail at Mid and End Placement Stages of the Assessment

Statement rated by Clinical Educator / *Student*	Response category and median rating	Clinical Educator (% responding)	Student (% responding)
Having more detail in the form of ratings on elements at mid placement helped me with my teaching/*learning*.	Agreed	73.5	85.2
	Neutral	16.2	6.8
	Disagreed	10.3	8.0
Clinical Educator N = 68 Student N = 88	Median rating	5	6
The student(s)/*Students* should also be rated on elements at end placement.	Agreed	42.6	42.5
	Neutral	11.8	21.8
	Disagreed	45.6	35.6
Clinical Educator N = 68 Student N = 87	Median rating	4	4
There was too much detail at mid placement.	Agreed	22.7	19.3
	Neutral	7.6	20.5
	Disagreed	69.7	60.2
Clinical Educator N = 66 Student N = 88	Median rating	3	3

8.3.4. Generic Competencies

The Generic Competencies are proposed to underpin or enable the competent practice of the occupationally specific competencies outlined in the CBOS and their inclusion was strongly supported by respondents (Table 54). Positive ratings for Generic Competencies were correlated with more experience for CEs (Table 48), although not for students who presumably do not have sufficient experience or knowledge of the practice of the profession for this to influence their ratings.

Both CEs and students were divided as to whether the Generic Competencies were more useful in the assessment than the CBOS competencies, with comments by CEs indicating that this was probably due to both sets of competencies being considered equally important. Three students commented specifically and very positively on the inclusion of the Generic Competencies in response to the final open ended question regarding feedback on any aspect of the tool e.g. *"Generic Competencies – makes the assessment much more thorough and gives a better view of where I am as far as overall learning is concerned."* (Student feedback 1).

The availability of the full version of the Generic Competencies was rated as being useful by 73.1% (N= 30) of the 41 CEs who responded to this question regarding the materials available in the assessment resource manual, and 67.6% (N=53) of the 75 students who responded. This supported the inclusion of this resource in the manual.

Table 53. Clinical Educators' and Students' Feedback Ratings in Response to Statements Regarding the Generic Competencies

Statement rated by Clinical Educator / *Student*	Response category and median rating	Clinical Educator (% responding)	Student (% responding)
The Generic Competencies assisted my judgement of my student's competence.	Agreed	85.9	85.4
	Neutral	9.4	11.0
	Disagreed	4.7	3.7
Clinical Educator N = 64 Student N = 82	Median rating	6	6
The Generic Competencies were an unnecessary inclusion in the assessment.	Agreed	3.1	3.7
	Neutral	6.2	8.5
	Disagreed	90.8	87.8
Clinical Educator N = 65 Student N = 82	Median rating	2	2
The Generic Competencies reflect knowledge, skills, and attitudes of value to the profession.	Agreed	89.2	91.5
	Neutral	10.8	4.9
	Disagreed	0	3.6
Clinical Educator N = 65 Student N = 81	Median rating	6	6
The Generic Competencies were good descriptors of the competencies that underpin competent practice of speech pathology.	Agreed	89.4	89.1
	Neutral	9.1	8.5
	Disagreed	1.5	2.4
Clinical Educator N = 66 Student N = 82	Median rating	6	6
The CBOS competencies are more helpful than the Generic Competencies when assessing a student's competency.	Agreed	22.7	24.3
	Neutral	34.8	46.4
	Disagreed	42.4	29.3
Clinical Educator N = 66 Student N = 82	Median rating	4	4

8.3.5. Validity

Overall respondents were positive about the assessment tool's effectiveness in supporting their judgement of competency, identification of students' strengths and weaknesses, and goal

setting, although students were more likely to rate the latter neutrally (Table 55). This may be due to the usual university assessment tool being used for goal setting. Only 13.4% (9) of CEs and 14% (12) of students indicated that they did not prefer the research assessment tool over other assessment tools, which implies that the large majority of respondents felt that the research tool was at least as good as or better than current tools. CEs' perceptions of the tool's validity were not influenced by their experience, which suggests the tool has strong face validity for all CEs. However more experienced students did feel it more effectively supported their judgement of their competency and identified their strengths and weaknesses, which may reflect their developing understanding of the learning task (Table 49).

Table 54. Clinical Educators' and Students' Feedback Ratings in Response to Validity Statements

Statement rated by Clinical Educator / *Student*	Response category and median rating	Clinical Educator (% responding)	Student (% responding)
The research Assessment Tool did not effectively support my judgement of the student's / *my* competency.	Agreed	10.3	10.5
	Neutral	2.9	7.0
	Disagreed	86.8	82.5
Clinical Educator N = 68 Student N = 86	Median rating	2	2
Overall, the research Assessment Tool is preferable to other clinical performance assessments I have used / *other clinical performance assessments clinical educators have used with me.*	Agreed	68.7	70.9
	Neutral	17.9	15.1
	Disagreed	13.4	14
Clinical Educator N = 67 Student N = 86	Median rating	5	5
The research Assessment Tool is effective in identifying the student's / *my* strengths and weaknesses.	Agreed	86.8	86
	Neutral	8.8	9.3
	Disagreed	4.4	4.7
Clinical Educator N = 68 Student N = 86	Median rating	6	5
The research Assessment Tool did not help in goal setting with the student(s) /*did not help me goal set.*	Agreed	11.8	9.6
	Neutral	10.3	33.3
	Disagreed	77.9	57.2
Clinical Educator N = 68 Student N = 63	Median rating	2	3

8.3.6. Electronic Version

The questionnaire included questions regarding the use of the online and hard copy versions. Only responses from CEs are described here as students were only given access to

their demographic data online as research funds did not extend to providing them with a parallel online version of the assessment tool. Thirty-eight CEs indicated that they looked at or used the online version of the assessment tool (57.6%).

Those who responded that they did not use the online version (N = 28) were asked for their reasons why and what might encourage them to use an online assessment tool. Twelve nominated that they did not have access to a computer connected to the internet and 6 had difficulty accessing the research website. Thirteen CEs stated that they preferred to work on a paper version or had to complete the assessment offsite. Only 2 CEs did not feel confident using the online version. Nineteen CEs responded that they might use an online version if they had better access to a computer connected to the internet. The need for training and support was nominated by two CEs. Familiarity with the tool appeared to be an important issue as11 of the 28 CEs asserted they might use an online version if they were more familiar with the assessment.

For those who did use the online version, 27 disagreed that the assessment tool was difficult to use (N=32) with only 5 agreeing. The majority of CEs agreed that assessments should be available online in the future (33 of 37 CEs who had used the online version), with 2 being neutral and 2 disagreeing with this statement.

This feedback suggested that those who used the online version of the assessment tool were positive about its ease of use and would like to have an assessment tool available online in the future. Very few CEs indicated that they did not use the assessment tool due to lack of confidence, training, or support, signifying that CEs are comfortable with computer delivered assessment formats. Most CEs cited access or familiarity issues, with some preferring to use a paper version, possibly for both these reasons. One CE mentioned the problem of providing privacy for a student assessment with computers in shared office space.

Some specific comments were made regarding improving the utility of the online assessment tool, which will usefully inform any future developments in this area. Five CEs commented very positively about the online option at the end of the questionnaire, e.g., *"The online version was fantastic!"* (CE feedback 22); *"Loved online!"* (CE feedback 6).

8.3.7. Marginal Students

Students and CEs were asked to state if they or their student(s) was considered to be at risk of failing the placement or their performance considered to be of concern at any time.

Those who answered ‘yes’ to this question were asked to respond yes/no to three follow up questions regarding how well the tool assisted in discriminating between marginal/acceptable performance, if the tool enabled problem areas to be specifically identified, and if the Generic Competencies were helpful in this process. Four students identified that they had been considered at risk. Two of the 4 students did not provide an ID to crosscheck against their ratings, of the 2 students who did, one had 3 units at mid placement identified as ‘of concern’ and the other had none. As none of the 4 students answered the three follow up questions, no further analysis of their responses to the feedback questionnaire was made.

Seventeen CEs indicated that their student(s) had been considered at risk, the majority of these CE’s answered the subsequent three questions. Thirteen CEs agreed that the research tool assisted in discriminating between marginal and satisfactory performances, with two disagreeing. Comments highlighted that the research tool assisted because it identified that students were performing differently to their peers on placement. CEs feedback identified that the examples, presumably on interpreting the behavioural descriptors in relation to the units of assessment, were particularly useful and gave clearer ‘expectations’ than the current assessment tool. One CE commented that *“Being able to see if student had progressed to entry level was made easy by being able to compare with intermediate and novice on the same page for each area”* (CE feedback 41). Two CEs commented on the specificity of the tool:

> *“Specific enough to identify why the student wasn’t performing at a higher level”* (CE Feedback 49)
>
> *“The increased specificity of the applied behavioural descriptors helped me to ask where change is needed.”* (CE feedback 51)

Fifteen CEs felt the assessment tool enabled the problem areas to be specifically identified, with two disagreeing. Four comments identified the usefulness of the mid placement assessment for articulating problem areas and goal setting, with 2 CEs commenting that the research tool assisted in isolating the difficulties the student was having. Fourteen CEs agreed that they found the generic competencies helpful in the process of managing a marginal student, with three disagreeing. Few specific comments made on this aspect of the research tool.

Overall the responses suggest that the CEs were very positive about the effectiveness of the tool in discriminating between marginal and acceptable performance and its ability to identify problem areas and found the generic competencies helpful. Interestingly, of the 17 CEs who indicated on their feedback that their student’s performance had been considered to

be of concern, only 8 had identified on the assessment tool at mid or end placement assessment that their student's performance on a particular unit placed them at risk of failing their work placement. This implies that CEs were more prepared to answer "yes" to the more generally phrased concern in the questionnaire (Was the student at risk of failing the placement or their performance considered to be of concern at any time?) than the more direct question in the assessment tool (Does performance on this Unit place the student at risk of failing this clinical experience?). Experimenting with providing less decisive statements about students' performance on the assessment on future versions of the assessment tool may be warranted to ensure that at risk students are identified early and assisted appropriately.

In general there was a trend for CEs who had marginal students to be more positive about all aspects of the research tool. Mann Whitney U's comparing rating patterns on 3 questions yielded statistically significantly different results between CEs who had marginal students than other CEs. CEs who had marginal students were significantly more likely to indicate that their teaching was helped by having more detailed ratings at mid placement ($p = .017$), that students should be rated on elements as well as units as end placement ($p = .017$), and that the research assessment tool assisted with goal setting ($p = .021$). Not surprisingly, Chi Square calculations demonstrated that CEs who had marginal students were significantly more likely to refer to the element level of the tool at end assessment than other CEs ($p = .04$). Overall, this confirms that having layers of detail within the assessment tool that can be accessed when determining students' competence was a useful aspect of the tool design.

8.3.8. Overall Satisfaction

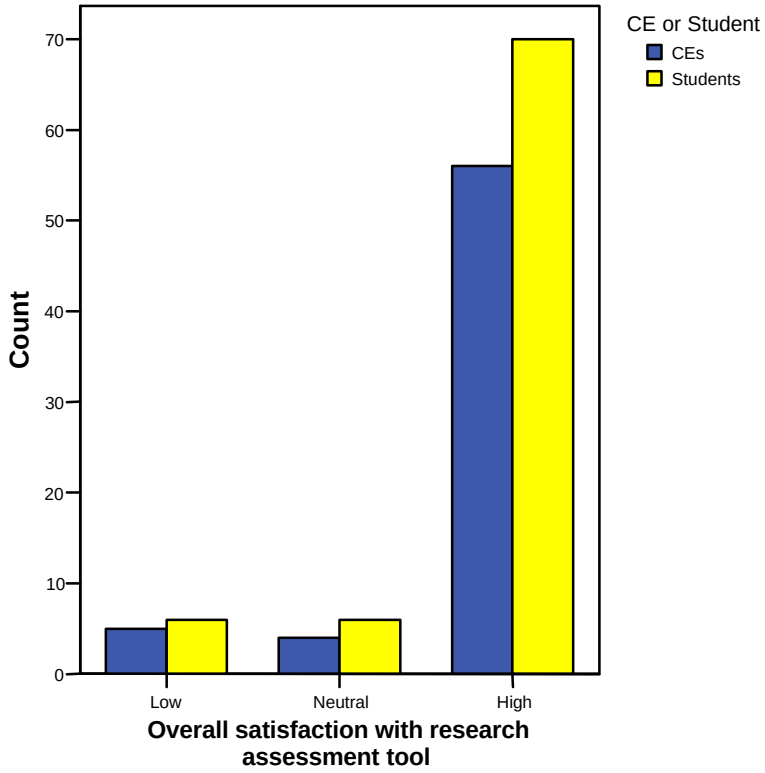
CEs and Students reported high levels of satisfaction with the research assessment tool, with the median rating for students being slightly higher (6) than for CEs (5.5) (Table 56). Satisfaction was not related to experience for CEs, confirming that face validity was strong for all CEs. However, more experienced students were significantly more likely to indicate a higher degree of overall satisfaction with the assessment tool (Table 49), again suggesting that students' understanding of the learning task may develop with time. In general, students provided a high satisfaction rating regarding the overall validity of the assessment tool and were significantly more likely to rate virtually all aspects of the tool positively (23 out of 27 statements rated). CEs showed a similar pattern, with 17 of 27 statements correlating significantly with overall ratings of satisfaction and to statements linked to all aspects of the

tool: rating scale; behavioural descriptors; detailed mid placement assessment; resources provided; generic competencies; and validity statements.

Table 55. Clinical Educators' and Students' Feedback Ratings Regarding Overall Satisfaction With the Assessment Tool

Statement rated by clinical educator / *student*	Response category and median rating	Clinical Educator (% responding)	Student (% responding)
How would you rate your OVERALL satisfaction with the research Assessment Tool as an assessment of your competency in your clinical placement?	High	86.2	85.4
	Neutral	6.2	7.3
	Low	7.7	7.3
NB ratings were from Low (1) to High (7)			
Clinical Educator N = 65 Student N = 86	Median rating	5.5	6

Figure 26. Overall satisfaction with assessment tool for clinical educators and students



8.3.9. Other Comments

In addition to the comments described above, a few CEs commented regarding the inclusion of a comments section under each unit in the mid and end assessment. Five found

that it was too time consuming and/or repetitive to comment for each unit, with one CE stating that she liked this aspect of the tool.

8.4. Summary

The prototype assessment tool was very well received by both CEs and students. Most respondents preferred it to current tools indicating that the innovations regarding scale design, assessment process, and inclusion of Generic Competencies were seen as appropriate and useful and therefore should be retained. The assessment's positive reception also suggests that face validity was strong and will facilitate its reliable and valid use as an assessment and teaching tool. Experience was significantly related to confidence in using the new tool as well as approval of some aspects of its design and content. Thus, investigating the effect of training and familiarity on user satisfaction with the tool will be an important line of future enquiry.

DISCUSSION

CHAPTER NINE

9. DISCUSSION

This thesis has outlined the rationale for and process of developing an assessment of speech pathology students' competency in the workplace, as well as the procedures undertaken to collect evidence regarding the assessment's validity. The key question is whether the evidence supports the assertion that the assessment tool can be validly used for the purpose for which it was designed. As argued in the literature review (Section 3.3.1.), the most appropriate criteria against which to evaluate the success of a performance assessment are those developed by Messick (1989; 1994; 1996). The following section will compare and discuss the evidence for this assessment tool's validity against these six validity criteria, identifying limitations, and possible future research directions in an endeavour to provide a "broader and richer" approach to validation of this research, moving "beyond the worn-out notions of content and predictive validity" (pp. 257, McGaghie, 1993).

9.1. Content Validity

This thesis has described the process used in this research to ensure that that the content of the assessment tool was relevant and representative of the construct domain of entry-level speech pathology competence and demonstrated a high level of technical quality as required by Messick (1996). Ensuring that the construct was neither over represented through inclusion of irrelevant assessment tasks nor under represented through omission of relevant aspects of speech pathology entry level competency was a critical component of this process.

Assessment design included investigating theory and recommended practice regarding the nature and role of judgement in assessment; and supporting that judgement to safeguard validity. These aspects were incorporated into the assessment design and Phase 1 described the technical aspects of the tool design intended to support judgement. This included attention to wording, judgement criteria (behavioural descriptors) that focus on qualities of performance, including a formative and summative component to the assessment, and providing supporting layers of detail and information the assessor could use to uphold the process of judgement.

The comprehensive developmental process undertaken ensured that the construct was neither under or over represented. Theoretical analysis of the nature of competency and how

this relates to the specific domain of speech pathology practice was undertaken. The construct boundaries to be assessed were specified and the attributes that the assessment was to reveal were described (see Phase 1). The theoretical understanding of speech pathology competence was integrated with a variety of other sources including current practice and documents such as the CBOS, and the practical and personal knowledge of competency held by students, experts in clinical education of speech pathology students, and practising CEs based in workplaces and university programs. In the final assessment tool these attributes were conceptualised as a combination of generic and occupational competencies that would be evaluated across the range of speech pathology practices represented by the particular workplace placement and were detailed in the assessment materials.

The validity of the newly developed generic competencies was judged to be high by both the CEs and students who were involved in trialling the assessment tool. This was reflected in their strong support for the inclusion of the generic competencies in the assessment and their agreement that these competencies were relevant, that they assisted in judgement of competency and that they were necessary for the competent practice of speech pathology. User feedback also highlighted that both the occupational competencies (CBOS) and generic competencies were seen as equally important in the assessment.

Attention was also paid to the development of performance standards for each item of competency. Examples illustrative of various levels of performance to reflect increases in the complexity of the construct of competence were developed. Both were undertaken with a view to minimising contamination of ratings by other sources of construct irrelevant difficulty. The high Person Reliability statistic of .98 (Section 7.5.2.) indicates a strong level of internal consistency within the assessment (Stone, 2003). In addition, person measures ranged from – 14.2 to 13.1 (Section 7.3.2). Both these statistics indicated that a large spread of ability and thus a clear hierarchy of development on the construct was identified by the assessment. This also reinforces that construct irrelevant variance was kept to a minimum.

9.2. Substantive Validity

Substantive validity is closely related to content validity but focuses more on identifying evidence that the assessment tasks are appropriately sampling the domain of competence and empirical evidence that the proposed theoretical processes are actually engaged in by the assessees during the assessment process (Messick, 1996).

9.2.1. Workplace Based Assessment

Substantive validity of the assessment tool is strongly promoted by the format of the assessment as it requires CEs to identify the competencies being engaged in by students as part of carrying out everyday speech pathology workplace tasks and to judge the level of competency illustrated by students' performance of these tasks. The decision to situate the assessment in the context of the real world arena of professional practice avoided construct over or under representation and meant that the assessment did not have to rely on unproven links between decontextualised assessment tasks such as OSCES, case presentations, computer or model based simulations to name a few, and competency in the workplace.

In addition, workplace based assessment ensured that the tasks undertaken by students as part of the assessment were highly meaningful and authentic to the competent practice of speech pathology. The decision to carry out the assessment in the workplace by the person who knows the student(s) most thoroughly ensures that the assessment process is perceived as valid by both parties and this was confirmed by user feedback strongly supporting that the assessment process and context was highly meaningful. This process ensures the performance assessment includes all the important factors that interplay in real life performance (Friedman & Mennin, 1991) as students will need to demonstrate that they can manage factors that may prevent their potential competence from being demonstrated in their performance such as their health, stress, and workplace factors (Rethans et al., 2002).

9.2.2. Effect of Experience

There was some qualitative evidence, not unexpectedly, that students' ability to achieve competency in a particular workplace task would vary according to their opportunity to engage in these activities (Section 7.4.1.). This was particularly evident with regard to the CBOS Competency 2 Analysis and Interpretation, and CBOS Competency 1 Assessment.

These competencies had item difficulties indicating that they were respectively the first and second most difficult items on which to achieve competency (Section 7.4.5.) and respectively had the least and second to least number of ratings made on end assessment. This suggests that not only are these items more difficult to gain competency in, relatively fewer students were having an opportunity to develop these competencies in the workplace. As the item difficulties are calculated on the basis of the ratings that were provided, having fewer ratings is not likely to have substantially affected the item difficulties calculated.

However, if students are gaining less experience in these competencies it can be expected that developing competency will be more difficult. Interestingly, the order of item difficulties (Section 7.4.5.) related well to item difficulties resulting from a Rasch analysis of a workplace based rating format to assess physical therapy students' competency. This research found the competencies Communication, Professionalism and Treatment were easiest to acquire and Knowledge and Problem Solving (which underlie CBOS Competencies 1 & 2) the most difficult (Rheault & Coulson, 1991). Whether this difficulty in acquiring these competencies is an artefact of experience or intrinsic to the nature of these competencies requires further research.

9.2.3. Generic Components of Competency

High intraclass correlations between person scores for the same student in two different but concurrent placement types (Section 7.5.4.2.), and in some cases with different clients, suggested that the competency levels the students were achieving may have been transferring across the range of speech pathology practice. This suggests that there are generic components to competency as originally proposed and students are engaging in the construct as expected. The way in which competency is transferred to and interacts with different knowledge bases or service delivery models of speech pathology practice and experience is worthy of future attention. Aspects of this issue will be discussed further in relation to consequential validity.

9.2.4. Developmental Progression in Competency

The process of achieving a pre entry level of competency in speech pathology on this assessment was conceptualised as being developed along a continuum representing novice through to entry level performances. The fact that the Rasch assumption of unidimensionality was upheld through strong item fit statistics (Section 7.4.2.) confirmed that students with low

levels of competency tended to score lower on the assessment items (competencies) than students with high levels of competency. This unidimensionality indicated that the students' assessment results were the result of their ability in combination with the relative difficulties of items and is strong evidence that the theoretical processes proposed are in fact being engaged (Fisher, 1994).

Another source of evidence supporting the substantive validity of the assessment is the way in which increasing levels of performance (person scores) related to increasing levels of experience (Section 7.5.4., Figs. 19 & 20, Table 38). Thus person scores were demonstrated to increase both qualitatively and in statistically significant ways over time within the small sample of students who had assessment information collected longitudinally. Also, scores increased significantly across groups of increasing hours of experience and students in final assessments scored significantly higher than students in their first placement. However, hours alone were not the sole predictor of higher levels of competence suggesting that other factors are instrumental in the development of competence.

Some evidence was found that indicated that the developmental progression towards entry level competency varied according to the university program attended although the differences were not present for students of higher ability (Section 7.5.4.2., Table 40). The effect of curriculum and other potential factors, such as the specific nature of student experience within different workplace practica, upon the development of competence promises to be an interesting line of future inquiry.

9.2.5. Marginal Students

Theoretical assumptions about the inherent variability of the performance of students who were judged as having difficulty in developing competence in speech pathology practice was also upheld, with the majority having higher than acceptable IMS values (above 2.0) (Section 7.5.3.3, Figure 18). Evidence was also found that students who were in the intermediate stage of developing competency were also more variable in their performance than students approaching entry level (Section 7.5.3.3., Figure 17). However, novice students may also be particularly variable in their performance, as the first two categories (describing novice to intermediate performance) on the VAS were larger than subsequent categories (7.2.1.), suggesting that there was considerable variability in the ratings given to this group. Further research is required to identify if greater variability is related to substantive validity or is an artefact of the greater range of rating categories for intermediate students than for others.

9.2.6. Making Judgements

Global ratings of the students' overall performance by CEs were strongly correlated with each of the item ratings (Section 7.5.4.2., Fig.23). However, there was not a complete one to one correspondence between the zone of competency indicated by the category represented by the Rasch person score and the category indicated by measuring the overall global rating on the VAS. Correlations between the global rating and individual item ratings suggested that some items were more influential than others in the CEs' assessment of the students' competency (Section 7.5.4.2, Table 41). This is not necessarily a problem for the validity of the tool particularly given the strength of all item to global rating correlations and the demonstrated unidimensionality of the tool. However, it does suggest that the construct of competency may be composed of subcomponents of varying relevance to CEs' process of judgement.

Less experienced CEs' judgement may have differed from more experienced CEs on four competencies (Section 7.5.4.2., Table 39). However, as sample size for the least experienced group was particularly small it is difficult to know if these differences were real. In addition these differences appeared to resolve for students with competence levels approaching entry level suggesting that all CEs, regardless of their levels of experience, have a clear idea as to what workplace performances constitute entry level competence. Further research would be useful regarding the effect of CEs' experience on the ability to judge competence

The unidimensionality evidenced by the Rasch analysis (Section 7.4.2.) indicates that generally CEs engaged in the rating task in the predicted manner, confirming that it related well to their understanding of competency and how it developed. There was some evidence that a few CEs did not use the rating scale in a way that was not congruent with the way in which the development of competency was conceptualised (Section 7.5.3.3). Some had a number of assessments with overfitting or underfitting patterns of ratings and it seemed, when examining ratings of entry level students who had person measures below the cut off point of 10.11, that some CEs may have idiosyncratic patterns of ratings. However it was not possible to confirm this without further investigation. Use of Rasch analysis illustrated that the person scores and IMS results yielded by the tool can alert assessment coordinators to occasions when the substantive validity of the tool is being negatively affected (7.5.3.). This enables appropriate action to be taken to investigate and remediate this situation e.g. training the CE, reviewing the evidence in support of a pass/fail decision.

9.2.7. Inclusion of Generic and Occupational Competencies

The fact that the item fit statistics provided by the Rasch analysis fell within the conservative range of 0.8 and 1.2 (Section 7.4.1.) was strong empirical justification for inclusion of all items in the assessment as it was clear that performance on the items was not being influenced by other construct irrelevant factors. Thus each item in the assessment contributed to the measurement of the construct of entry level speech pathology competence. This confirmed the proposed theoretical construct that both generic and occupational competencies, and their related knowledges, skills, and personal qualities, are indeed integral to appropriate professional practice in speech pathology. The strong item fit statistics also mean that it can be justifiably assumed that the assessment tool assesses an underlying unidimensional trait and reliably places students along the continuum of this underlying trait.

The finding that generic competencies are intertwined and as important as occupational competencies may have implications for the content and processes of speech pathology educational programs and future revisions of the CBOS. The related lifelong learning skills of reflection and self evaluation are frequently addressed in the professional preparation component of curricula and this practice is recommended by speech pathology educators (Robertson et al., 1997) as well as others who promote the role of lifelong learning in ensuring ongoing competence (Boud, 2000). Skills such as clinical reasoning are already explicitly taught and facilitated within some speech pathology programs (McAllister & Rose, 2000) and no doubt other generic competencies are as well. Given that the Generic Competencies included in the research assessment tool have been found to be an allied and important component of speech pathology competency, it may be useful to ensure that they are explicitly attended to throughout the curriculum.

Finally, the positive feedback from students and CEs on their use of the assessment tool suggests that both parties were likely to fully engage with the assessment tool and the constructs involved (Chapter Eight). This engagement will promote its valid use as an assessment and learning tool.

9.3. Structural Validity

A considerable amount of attention was paid to the fidelity of the scoring structure to the structure of the construct domain of speech pathology competency and is a unique aspect of this research that greatly strengthens the validity of the assessment tool. As recommended by Clauser (2000) and described in Chapters Four and Five, the research addressed scoring issues such as the aspects of performance to be scored, the criteria to be applied to produce a score, and how this was developed; these aspects are covered in the prior sections on content and substantive validity. The structural validity evidence rests on how the performances would be rated (Clauser, 2000) as well as how this relates to the construct domain being assessed.

9.3.1. Generating a Score

The decision to situate the assessment in the workplace and use the CE(s) who is in regular contact with the student(s) rather than a visiting evaluator, ensured that multiple, regular, and detailed observations underpinned the judgements involved; and thus underpinned the final score for each student. This is similar to the multiple real world observation approach to assessment now being advocated in the medical field and suggested to improve the validity of performance assessment e.g. Turnball, McFadyen, van Barneveld & Norman (2000), Page (2004), Cox (2000). However, it is important to note that this requires a person who has established a close working and teaching relationship with the student(s), i.e. the CE, to make a potentially high stakes judgement that could have significant negative impact upon the student(s). Authors such as Duke (1996) and Ilott & Murphy (1997) have highlighted that this can create significant ethical and moral conflict for CEs and may affect their judgement, particularly in relation to making the decision to fail a student in a placement.

9.3.2. Evaluating the Rating Scale

In addition, it was considered fundamental to the validity of the assessment tool that the ratings reliably identify discernable degrees of competency in an unambiguous and ordinal fashion i.e. that a rating of 2 was highly likely to reflect a lower degree of competence than a rating of 3. Without this, it would be impossible to construct meaningful measures of person ability as any analysis requires data to be at least ordinal. Lopez (1996) terms this property

‘communication validity’ in that it identifies whether the raters “converse with the test developer in a common language free of idiosyncratic category usage, response sets, and ambiguous terminology”(pp. 482). This communication validity is established through ensuring the rating scale categories perform as intended, that respondents could discriminate and order the response levels involved. In fact, Lopez asserts that it is “pointless to examine any other form of validity until we have established that we have listened carefully to what test respondents have told us about our variable” (pp. 482).

The process of evaluating if the rating scale gathered data that could be usefully converted into a measure of competency was an iterative process (see Chapters Four and Five). The 100 categories of data represented by VAS ratings were grouped and regrouped until the Rasch analysis indicated the greatest number of well functioning and clearly identifiable ordered categories existing in the data. These categories represented the number of reliable discriminations of levels of competency that CEs were demonstrating in their use of the VAS. The analysis indicated that judgements of competence represented by VAS markings could be confidently segmented into 7 categories of different sizes: 0 to 25, 26 to 50, 51 to 62, 63 to 74, 75 to 87, 88 to 99, and 100 and above entry level (Section 7.2.3.). This finding of uneven intervals and less than 10 categories along the VAS was similar to that found in other research on the use of VAS (Cook et al., 2001; Munshi, 1990; Powell, Kelly, & Williams, 2001; Thomee et al., 1995) although it is the first time it has been found in relation to rating performance.

This categorisation allowed confidence that ratings in each successive category reflect meaningful distinctions between levels of competence in speech pathology practice and indeed translates the students’ performance into a score that adequately represents their performance (Clauser, 2000). It also ensured that the ratings defined the continuum of competence and identified the maximum number of categories that could be used meaningfully and thus maximise the precision of the assessment tool (Andrich, 1999).

The Rasch analysis of the rating scale provided strong justification for the scoring procedures used in the assessment format (Clauser, 2000). It is recommended that the rating task continue to take the format of a VAS scale without category markings as the validity of the scale may rest on the nature of the assessment task it presented to CEs. It is possible, as suggested by Averbuch & Katzper (2004), that respondents using a VAS scale are likely to feel more comfortable with the decision to move a mark slightly on the continuum as it does

not confront them with the decision as to whether the difference perceived actually justifies a change in a whole category up or down a categorical rating scale.

The Rasch analysis of the VAS strongly contributed to the validity of the tool as it ensured that the ordinal nature of the ratings and the degree to which CEs could discriminate competency was not assumed but in fact demonstrated. Thus the scoring model developed is based on what was discovered through analysis regarding the way in which CEs mapped the construct of developing competency along the VAS rather than imposed by the researcher's preconceptions. In addition, careful calibration of the assessment tool (Section 7.3.) ensured that only ratings that conformed closely to the Rasch model had acceptable IMS values. Requiring this stricter level of measurement ensures that scoring is more exact and thus any scores that are at all questionable, in terms of how well they fit the Rasch model and provide an accurate measure of students' competency, will be identified through IMS values of above 2.0.

9.3.3. Other Considerations

The use of Rasch analysis also results in a person score with an estimated error variance for each person rather than the sample overall. This degree of exactness further safeguards the validity of the scoring process. In addition scores are highly likely to be comparable across groups as they are sample independent due to the process used by a Rasch approach. This aspect will be described further under the generalisability aspect of validity.

Finally, the structural validity of using two assessment formats, one in hard copy and one electronic, was evaluated and it was found that there were no overall significant differences related to the rating format (7.4.4., Table 32). Thus it would appear that the physical format, e.g. VAS length, did not have a significant effect on the structural validity of the tool and perhaps the cognitive task of transforming a judgement onto a mark on the VAS is the key aspect of the assessment, not what kind of mark on what kind of line. This is interesting given exhortations in the literature to ensure that the same line length is strictly adhered to (Johnson, 1997) and confirms Ahearn's (1997) suggestion that there is no scientific reason for the length chosen.

9.4. Generalisability

This aspect of validity is concerned with the extent to which score properties and interpretations generalise to and across population groups, settings, and tasks (Messick, 1996). As a first step the assessment was designed to be highly representative of the content and processes of the construct domain. Thus the person measure (or test score) can be interpreted with confidence as representing an actual degree of competence in that workplace. The demonstrated increase in this measure over time and with experience (Section 7.5.4.2., Figs. 19, 20 & 21) coupled with similar levels of performance across different placements over the same time frame (Section 7.5.4.2.) suggests that it can be assumed that the measure represents a level of competence that is generalisable across the scope of practice of speech pathology. Further investigation is required to confirm this and to identify any limitations that may exist when generalising person measures from one workplace to another.

9.4.1. Generalising Across Groups

Rasch analysis also ensures that scores are highly likely to be comparable across groups as the analysis estimates the difficulty of assessment items independently of the sample used and uses an algorithm to describe an item characteristic curve (see Fig. 14, Section 7.4.4. for an example) which enables prediction of how a person of any level of ability is likely to perform on each item (Clark & Watson, 1995). Thus, if the data fits the assumptions of the model, Rasch analysis is able to calibrate items in an assessment over the complete range of possible scores, even if the Calibration Sample does not represent all possible scores. Naturally, larger and more heterogenous the sample used for calibration of the scale and items lead to more accurate estimations i.e. with a smaller degree of variation either side of the item characteristic curve. In the case of this research, a heterogenous sample was collected with regard to levels and types of student experience. However, the majority of the sample was collected from students at The University of Sydney. While this is unlikely to have a severe effect on the test calibration, it would be useful to confirm the item difficulties and rating scale categories on a sample representing a broader range of speech pathology programs.

Two statistics are generated by the Rasch analysis that provide information on how generalisable the assessment is likely to be: Item and Person Reliability indices (also called separation indices, Section 7.4.3 & 7.5.2.). These statistics are described as being analogous to Cronbach's alpha (Bond & Fox, 2001). The closer the value of these indices to 1.0, the

more likely that the variations in person ability are due to actual differences rather than error (Andrich & Sheridan, 2004b). The Item Reliability was measured as .97 indicating that the items were stable i.e. likely to represent the same level of difficulty even if they were used to rate another group of students with similar ability levels (Bond & Fox, 2001). The Person Reliability index was .98 and indicated that, if this group was given another set of items that measured the same construct, each person was highly likely to keep their place in the order of least to most competent (Bond & Fox, 2001). This compares favourably with a person reliability of .81 and item reliability of .85 found by Curtis and Denton (2002) in their workplace based assessment of problem solving and which they suggested indicated that their assessment tool had satisfactory measurement properties.

As already identified (Section 7.3.2), the large spread of person abilities represented in the sample indicates that the rating task is well targeted to the sample being assessed. Given the breadth of experience and placements represented in the sample, this further reinforces the strong generalisability demonstrated. Overall it can be confidently stated that the assessment can be reused with a similarly heterogenous group of speech pathology students and yield results that can be compared across samples.

9.4.2. Generalising Across Raters

Finally, high rates of inter rater reliability were found between the person measures yielded by ratings from two different CEs for the same student at the same time (Section 7.5.4.2). The group of 16 students who were placed in a workplace with two different supervisors, and had person scores with IMS values below 2.0, yielded high intraclass correlation coefficients (ICC) of .87. Thirty-three students who had two CEs at two different workplaces and scores with IMS values below 2.0 also yielded a high ICC of .82. This degree of inter rater agreement is pleasing particularly considering that the assessment tool was relatively unfamiliar to the CEs (as compared to formats they used regularly) and they had no training in the use of the assessment tool with their students.

In general, the reliability of rating scales and/or raters is thought to be poor, particularly when used to assess performances in clinical settings (Dauphinee, 1995) and inter rater agreement of .80 or above is rarely reached. The research assessment tool compares favourably with reports of workplace ratings by Turnbull et al. (2000) for medical students. Intra class correlation coefficients for raters rating the same students were .00 to.22 for nursing supervisors rating on items on a multidisciplinary team form, .64 to .73 for items

rated by an attending faculty supervisor on an admission rating form, or .81 to .86 for supervisors' ratings on a ward rating form.

Cross et al. (2001) trialled two performance rating formats as well as comparing physiotherapy field CEs' overall rankings for six students' performance in video vignettes to university educators' rankings on each format. They found that university educators had high agreement regarding the rankings for the 6 students (Kendall's coefficient of .98) and that field CEs' rankings were moderately similar to university educators' rankings (Kendall's coefficient of .62) after training in the use of the most reliable format used in the research (Cross et al., 2001). However, individual field CE's correlations (after training) with the university educators' rankings on the most reliable form varied from .46 to .89 (Spearman's rho). Most reliabilities reported in the literature relate to OSCE style examinations and few of these are higher than .80. For example, correlations between ratings of examiner's pairs on 18 different patient stations on OSCE examinations of .38 to .91, with only 3 ratings correlating above .80 (Newble & Swanson, 1988). Cohen, Rothman, Poldre & Ross (1991) found ICCs of global (rather than behavioural checklists) ratings provided by three raters on 'approach' and 'attitude' of medical students on an OSCE patient station were .49 and .22 respectively.

The levels of rater reliability in this research compare favourably with research by Roach et al. (2002) who developed an assessment tool to assess the competency of physiotherapy students in the workplace. Their investigation informed this research greatly in terms of process and format, although a different approach to defining competency and criteria for rating was used. Intraclass correlations of .87 were found for joint evaluators of physiotherapy students and .77 for joint assessments of physiotherapy assistants (Roach et al., 2002).

The high degree of rating reliability for this research is likely to have been supported by the structural validity of the rating scale which ensured that the person measures were derived (through Rasch analysis) from meaningful rating categories. The achievement of strong reliability also confirms the assertion made in the literature review (Section 3.3.4.) that attention to the content, form, and process of assessment enables judgements of performance that are valid. Thus the assessment tool can be assumed to have adequately controlled for sources of error related to the judgement process. Overall, the process of investigating the structural validity of the tool and high reliabilities confirm Friedman & Mennin's (1991) suggestion that sampling many specific behaviours over time and in various situations may provide an approximation of the true performance. These consistencies give confidence that it

is not just a matter of personal opinion but that students are indeed demonstrating reliable behaviours (Friedman & Mennin, 1991).

The effects of training on improving inter rater reliability and thus further increasing confidence in the validity of the assessment tool, will be of interest. In addition, investigating whether raters are the primary source of variance or whether the ratings represent actual differences in performance in different contexts or with different clients would be a useful line of enquiry. However, it would be anticipated that variability in examinee performance would have the greatest impact on the score they receive rather than the rater's behaviour, particularly with a well designed assessment which has already demonstrated high inter rater reliability (Cohen et al., 1991; Keen et al., 2003; Newble & Swanson, 1988; Norman et al., 1991; Shavelson et al., 1993).

9.5. External Validity

External validity primarily rests on the investigation of convergent and divergent evidence about the assessment i.e. is it related in a logical and expected way to assessments that measure similar constructs or clearly unrelated to plausible alternative explanations for the score received on the assessment (Messick, 1996)? One could query the logic of relating the validity of a new tool to already existing tools on the basis that there is little point in developing a new one unless it either reduces measurement error or extends the measurement of the variable further into ranges not previously tapped (E. V. Smith, 2001). However, as there are no other validated tests of competence in speech pathology this is moot as there were no directly comparable assessments to correlate results with the research assessment tool. In addition, the scope and design of the research did not allow for taking other measures to investigate divergent or convergent validity. Given that generic competencies were included in the assessment, it may be that future research could investigate this aspect of validity further using assessment tools that are under development and related to generic competencies e.g. The Authentic Test of Problem Solving (Curtis & Denton, 2002) or the Graduate Skills Assessment (ACER, 2001).

One source of external evaluation of the assessment that was utilised was feedback from the CEs and students involved in the use of the assessment tool during the field trial. As described in Chapter Eight, the tool was well received by both groups and generally preferred to current tools indicating that the content and format was appropriate and should be retained. Experience levels of both CEs and students were significantly related to confidence in using

the new tool as well as approval of some aspects of its design and content and suggested that it had even stronger validity with those who understood the nature of the learning task involved and the judgements to be made.

9.6. Consequential Validity

It is critical to ensure that scores from assessments are interpreted in a justifiable manner before becoming the basis for action (Messick, 1996). While it is not possible to anticipate all potential uses and misuses of test scores, this does not release the test developer from the responsibility of considering these aspects of test use.

9.6.1. Effect of Assessment on Learning

The impact of the assessment tool upon learning was identified as a potential threat to consequential validity particularly given that competency based assessment has been criticised for negatively affecting learning (Wolf, 1995). Substantial effort was devoted to ensuring that the assessment tool had strong content validity such that the content directed students' attention, and indeed that of their CEs, to appropriate learning goals. It was also proposed that this was important for valid engagement in the assessment process by students and CEs (see Section 6.4.2.3, Chapter Six), further safeguarding consequential validity. A high degree of success was indicated by the strongly positive student and CE feedback regarding the content of the assessment (Chapter Eight).

Substantial benefits may accrue to teaching and learning through ongoing use of the features of Rasch analysis to evaluate the impact of different teaching and learning practices. For example DIF analysis was able to identify some differences in the acquisition of competence on particular items between students attending The University of Sydney and students from other universities (Section 7.5.4.2., Chapter Seven). This may yield useful information regarding the impact of different curricula upon the development of workplace competencies. In addition, the performances of cohorts of students across the different competencies can be tracked to identify whether there are areas of difficulty that need addressing prior to further placements or graduation.

Including a formative and summative assessment component also focuses attention on the learning aspect of assessment (see discussion in Section 5.4, Chapter 5) and was also well received by students and CEs. Strong positive feedback was made regarding the detailed

resource materials provided that would enable students to identify what learning was required and equip CEs to provide specific feedback and direction to students regarding their learning.

9.6.2. Interpreting Person Measures

The process of design and trialling the assessment tool has enabled the development of an assessment tool that can be used to derive a person measure plus individual margin of error for each student who is rated on this tool. This measure, or degree of competency demonstrated by students, can be used to confidently place students into one of seven zones of competency development, as indicated by the thresholds established during validation of the rating scale (Section 7.3.2.). Careful calibration procedures (as described in Section 7.3., Chapter Seven) and other factors discussed in the section on construct validity, indicate a high degree of confidence can be placed on how validly these scores describe students' competency levels.

Nevertheless, a confident interpretation of the person measure for students rests on the assumption that CEs share a common interpretation of the rating task. The high inter rater reliability scores indicate that this indeed did occur, despite the absence of training. However, the rating task introduces a major source of error that must be considered when interpreting the assessment score, particularly in high stakes testing situations such as determining whether a student passes or fails a placement or is sufficiently competent to graduate. The possibility that CEs may be using the rating scale idiosyncratically should be acknowledged and investigated. This is particularly important if CEs rate over a large spread of the VAS (as indicated by IMS values above 2.0) suggesting that a student's performance may be marginal and that the person measure not accurately reflect this student's level of competence. CEs who use only a small part of the continuum represented by the VAS, as indicated by low IMS values, may also need training in judging each competence independently from the other.

In addition, some students may be rated lower on particular competencies because of lack of opportunity to practice the competency either in previous or current work places. This lack of opportunity may result in lower competence and may not reflect their ability to develop competency in this particular area (as discussed previously). Further to this, it must be emphasised the assessment is related to students' performance in a specific workplace and not to the range indicators of the CBOS statement that specify the age groups and types of disorders across which these competencies must be demonstrated.

9.6.3. Predicting Future Performance

As discussed previously (Section 9.2.3.), competency does appear to transfer across placements to some degree but to what extent is yet to be determined. There is some evidence in the literature that while content specificity has been identified as an issue in OSCE assessments, i.e. that competency with one case does not guarantee competent performance on another, making global judgements of more generic components of performance may reduce this effect and may represent a broader and more stable aspect of performance than specific performance on a specific case (Govaerts et al., 2002; Keen et al., 2003). Thus speech pathology programs will need to give careful consideration to using the assessment tool to sample competencies across the range indicators required by SPAA and weigh up evidence of broader, more stable aspects of performance against levels of competency achieved with a client group representing a specific range indicator. Further research regarding the generalisability of competencies across models of service delivery and client groups, the role of the generic competencies in supporting this generalisation, and the relationship of entry-level competency to future workplace performance, will assist in the exercise of this judgment.

9.6.4. Determining Thresholds to Indicate Marginal Performance

The tool can potentially address one of the most difficult and critical aspects of clinical education: that of determining a threshold point to identify failing from non failing students (Hunt, 1992; Ilott & Murphy, 1997). First, with regard to entry level competency, the Rasch model identifies a threshold level to be the level at which a student has a 50% probability at being given a rating either side of the threshold. For example, a person score of 10.11 means that this student has a 50% probability of falling into the zone of competency (or category) 6 or 7, 7 being entry level. For other types of testing, where items become progressively more difficult, the Rasch model considers the 50% chance level as being the level at which the student's true competency lies i.e. if you have a 50/50 chance of getting the answer to a question correct, the question is probably targeted exactly at your ability level. In the context of rating competency, a score of 10.11or above indicates that it is probable that your overall performance across the items places your level of ability in zone of competency 7, or entry level competency.

The practical implications of this will need further consideration by the profession. In the research sample, students' person scores could clearly fall above the 10.11 threshold for entry level but still have 2 of the 11 competencies being rated as a 6, rather than 7. Category 6 covered measurements in the range of 88 to 99 as opposed to placing a mark on the end of the VAS scale at the 100 point or ticking the 'above entry level' box. Decisions regarding whether to routinely graduate students who are above the 10.11 cut off point but have not rated at 7 in all competencies (which would be represented by a maximum person score of 13.41) is a matter that will require consideration by speech pathology programs. Careful professional judgement by the education program will need to be exercised in relation to decisions regarding those students completing their final placements who have a person score that is marginal in relation to the 10.11 threshold or even below it. This will be discussed further below.

Second, high IMS values, indicating unusually variable ratings, were demonstrated to be a marker of marginal student performance across earlier levels of experience and a high IMS was combined with a low person measure for students considered to be marginal in their final placement (Section 7.5.3.). However, a high IMS value may also be a marker for students who have unexpectedly strong specific competency areas. In these cases, the students' ratings in relation to minimum ratings expected given the students' degree of experience/progression through the program would need to be taken into account when evaluating the meaning of the high IMS value. Those students who are marginal because their performances consistently fall short of the levels demonstrated by their cohort will be identified by a slower or different developmental pattern than their peers (Rubin, 1996) rather than a high IMS value.

This identification would ideally be supported by programs benchmarking the minimum level of performance required at the completion of each placement so it can be identified whether students are underperforming compared to their peers and whether this poor performance is maintained over more than one placement. There are a number of ways in which this could be done through tracking and comparing person measures or competence scores for each placement over the program. Expected performance levels or overall category of competence into which the students' competency score should fall could be identified for particular stages of the course. These performance levels could be matched with actual person measures achieved, the zone of competence these represent, and perceptions of CEs as to whether the students' performance met minimum expectations or exceeded them. In addition, over time person measures achieved by students deemed to have failed a placement could identify the minimum score required to pass a placement at a particular point in the program.

However, this may have to be restricted to those students who have scores with IMS values below 2.0 as higher values indicate that the score may not be an accurate measure of their competence.

Another aspect of making judgements regarding marginal levels of performance that is well supported by this assessment tool, are the aspects of competence identified and measured via the Generic Competencies. It has been identified that these generic aspects of performance frequently cause concern to CEs regarding students' competency (Altmaier et al., 1990; Hayes, Huber, Rogers, & Sanders, 1999). However, these concerns e.g. regarding professional behaviour, communication, or lifelong learning skills, are rarely translated into a failing grade probably due to concerns in appropriately documenting and measuring these behaviours (Carraccio et al., 2002; Hayes et al., 1999). The documentation and measurement of these types of competencies in the research assessment tool enables them to be an integral part of the assessment of competency in the workplace.

9.6.5. Accuracy of Measurement

Overall, it can be stated that the assessment will provide a valid person measure for each student on most occasions of its use as long as key aspects of its content, format, and procedure are adhered to, e.g. it cannot be assumed that the final assessment is valid if a mid assessment is not conducted as it may be that the mid assessment informs the final decision regarding competency. If speech pathology programs choose to undertake Rasch analysis of the ratings as well, an IMS value will provide further important evidence of students' competence and flag any person measures that may not be accurate. It is important to note that different Rasch analysis software use slightly different algorithms and provide different kinds of statistics and that consequential validity will be negatively affected if different programs, other than those used by this research (Winsteps or Bigsteps), are used without further interpretation of their meaning.

However, as mentioned above, construct irrelevancy may be introduced through idiosyncratic rating behaviour on the part of a CE or lack of opportunity to develop practical competency. Thus, if this evidence is used in a high stakes decision such as passing or failing a student, the meaning of the measures will need to be considered in consultation with the rater to make an informed judgement as to whether all the evidence supports the final decision and in the light of the student's learning opportunities and evidence of potential that the student will become and is likely to maintain competency into their professional future. To

ensure that the assessment is validly used, programs will need to consider how much weight the assessment information yielded by the tool will be given and how it will be factored into current and future protocols regarding assisting students who are having difficulty reaching passing standards for workplace competency and making fair and defensible decisions to fail students. Certainly the validity evidence is such that, once sources of construct irrelevant variance are considered, the research assessment tool yields strong evidence regarding students' workplace competence.

9.6.6. Integrating Scores With Other Measures

Integrating these scores with other, unvalidated sources of information regarding competencies related to performance but not assessed directly in the workplace, e.g. reflective portfolios to demonstrate lifelong learning skills, will require careful consideration. This is of particular concern when research has identified modest to low correlations between different types of decontextualised competency assessments (Edelstein et al., 2000; Newble & Swanson, 1988) or between assessments carried out in the workplace by a mentor and other forms of assessment of competency related to the workplace such as OSCES (Norman et al., 2002; O'Donohue & Wergin, 1978). With regard to speech pathology, a study by Begg and Ferguson (2004) found none to weak relationships between three types of assessment used to determine workplace competency (viva examination, portfolio, and CEs' ratings in the workplace) within the speech pathology program at the University of Newcastle.

There are a number of confounding factors that could create this lack of correlation between assessments including that the rating and scoring procedures used have not ensured that the data is truly ordinal or indeed interval, and thus can be usefully compared to each other. Indeed the assessments may not be sampling the same constructs and so results on each will not relate closely, however this research has identified that competency can be considered a unidimensional construct, or at least is the result of multiple psychological processes functioning in unison (Bejar, 1983, cited in Curtis, 2004). Therefore the research tool could form a strong starting point for relating evidence gathered from other sources using Rasch techniques for equating and linking procedures. This process uses persons and/or items that are in common across two or more assessments and would result in a common metric between assessments (Bond & Fox, 2001; Muraki, Hombo, & Lee, 2000).

The consequential validity of the assessment would be further supported if the assessment tool is part of an overall, integrated framework of evaluation that includes other sources of

evidence over the whole of a program. It is quite clear, for example, the assessment would not provide sufficient evidence regarding the soundness of the students' propositional knowledge base, though problems may be suggested by poor performances in workplace competencies related to clinical reasoning, analysis, and interpretation and planning. Developing this framework for collecting evidence across the program will also require consideration of the CBOS range indicators and how assessment of competency will relate to these. Examples in the literature include the assessment process undertaken to assess poorly performing medical practitioners in the United Kingdom (Southgate, Campbell et al., 2001; Southgate, Cox, David, Hatch et al., 2001; Southgate, Cox, David, Howes et al., 2001) or other frameworks suggested for fair and defensible assessment of practice (Lew et al., 2002; Schuwirth et al., 2002).

9.7. Summary

Developing a valid assessment of entry level speech pathology competency first required developing an understanding of the nature of speech pathology competencies and detailing these as well as the criteria against which to assess progress and achievement of entry level competence. The second phase involved undertaking appropriate validation procedures to evaluate whether this conceptualising of competency and its development was both appropriate and able to be validly assessed or indeed, quantified. As discussed, the research assessment tool has strong validity characteristics that will enable Australian speech pathology pre-professional preparation programs to use it with a high degree of confidence. A number of strategies for ensuring that this confidence is justified were identified, primarily through the use of Rasch analysis to analyse the assessment ratings provided by clinical educators. There are several consequential aspects of valid use of the assessment format that warrant careful consideration and discussion by faculty at speech pathology programs.

Limitations to the research impacting upon its consequential validity, were discussed and rested on two aspects of the research. First, conducting the assessment in the real workplace environment based on ratings of CEs was the source of some its greatest strengths e.g. construct representativeness, informed judgement, and direct links to real world performance. On the other hand it required judgement by those in close relationship with the student(s) and rested on use of a rating scale that could be used idiosyncratically by CEs when assessing. In addition, it was not clear whether the development of competency was intrinsically different for particular competencies or if this was due to the assessment occurring in a real workplace

environment where opportunities for learning and assessment are constrained by what is available in that context over that specific placement. The second limitation was the potential effect on calibration of the assessment tool due to the sample having a proportional over representation of students and CEs from one educational institution.

In the first instance it is recommended that a second version of the tool be created through revising the research tool on the basis of feedback from universities, students, and CEs in combination with the statistical analysis carried out. Further calibration should be carried out through involving a greater number of students from a range of university programs. Ideally, sources of invalidity such as rater variance should be further investigated as well as the effect of training and experience.

It would be useful to investigate further how the test scores should be incorporated into a framework of judgement regarding students' readiness to enter the profession given the limitations regarding valid interpretation of test scores and possible sources of invalidity identified in the discussion. This could include evaluating the possibility of equating and linking this assessment with other assessments of competence or performance. Suggestions made regarding use of test scores for benchmarking, identifying marginal students, and evaluating CEs could be considered when developing and testing this framework. In addition, further research on the assessment tool's accuracy and utility in identifying marginal students is important for ensuring appropriate judgements are made and timely remedial action taken.

Quality teaching and learning would be promoted through further research on the nature of competence, how it develops, and what aspects of it are most influential when making judgements about it. This could include investigating how competence transfers across client groups (range indicators), different service delivery models, and into the future when new professional learning is required. Investigating whether the item difficulties identified are intrinsic to the competency being developed or an artefact of relative amounts of experience would also be useful. Identifying what factors, other than experience, seem to be instrumental in developing competence will promote good educational practice such as addressing the role of the generic competencies in this process, or the type of placement, or quality of teaching provided by CEs. It may also be possible to evaluate the effect of different types of curriculum on the development of competency including how important it is to make generic competencies explicit within academic curriculum.

This research has developed the first prototype of a validated assessment of entry level speech pathology competence that is grounded in a unified theoretical conception of entry

level competence to the profession of speech pathology and the developmental progression required to reach this competence. This research will assist the profession of speech pathology in ensuring that speech pathologists enter the workplace well equipped to provide quality care to their future clients, the ultimate goal of any professional preparation program.

REFERENCES

ABIM. (1998). *Residents: Evaluating your clinical competence in internal medicine* [web article]. American Board of Internal Medicine. Retrieved 30 October, 2001, from the World Wide Web: www.abim.org/resources/publications/Resident.pdf

ACER. (2001). *Graduate Skills Assessment Summary Report: GSA Entry 2001*: Australian Council for Educational Research.

AERA. (1999a). Reliability and errors of measurement, *Standards for Educational and Psychological Testing* (pp. 25-36). Washington: American Educational Research Association.

AERA. (1999b). Validity, *Standards for Educational and Psychological Testing* (pp. 9-24). Washington: American Educational Research Association.

Ahearn, E. P. (1997). The use of visual analog scales in mood disorders: A critical review. *Journal of Psychiatric Research, 31*(5), 569-579.

Alexander, H. (1996). Physiotherapy student clinical education: The influence of subjective judgements on observational assessment. *Assessment & Evaluation in Higher Education, 21*(4), 357-366.

Altmaier, E. M., McGuinness, G., Wood, P., Ross, R. R., Bartley, J., & Smith, W. (1990). Defining successful performance among pediatric residents. *Pediatrics, 85*(2), 139-143.

Anderson, J. (1988). *The Supervisory Process in Speech-Language Pathology and Audiology*. Boston: College Hill.

Andrich, D. (1999). Rating scale analysis. In G. N. Masters & J. P. Keeves (Eds.), *Advances in measurement in educational research and assessment* (pp. 110 - 121). Oxford: Elsevier Science Pty Ltd.

Andrich, D., & Sheridan, B. (2003). RUMM 2020 [Statistical]. Perth, Australia: RUMM Laboratory.

Andrich, D., & Sheridan, B. (2004a). *Interpreting RUMM 2020 [Monograph Part II: Polytomous Data]*. Perth, Western Australia: Rumm Laboratory.

Andrich, D., & Sheridan, B. (2004b). *Interpreting RUMM 2020 [Monograph: Part I Dichotomous Data]*. Perth, Western Australia: RUMM Laboratory.

Andrich, D., & Wright, B. D. (1994). Rasch sensitivity and Thurstone insensitivity to graded responses. *Rasch Measurement Transactions [electronic journal], 2004*(16 July).

ANTA. (2002). *Learning and Assessment Strategies Part 2: Resource Guide*. Brisbane: Australian National Training Authority.

ANTA. (2003, 12 September 2002). *VET - What is it?* [web page]. Australian National Training Authority. Retrieved 20 June, 2003, from the World Wide Web: http://www.anta.gov.au/vetWhat.asp

AQFAB. (2002). *Australian Qualifications Framework: Implementation handbook*. Melbourne: Australian Qualifications Framework Advisory Board.

ASHA. (2000, 23/10/00). *Standards and Implementation for the Certificate of Clinical Competence in Speech-Language Pathology*. American Speech-Language-Hearing Association. Retrieved, from the World Wide Web: http://www.asha.org/about/membership-certification/handbooks/slp/slp_standards_new.htm

ATEAM. (2001). An ethics core curriculum for Australasian medical schools. *Medical Journal of Australia*(175), 205-210.

Averbuch, M., & Katzper, M. (2004). Assessment of visual analog versus categorical scales for measurement of osteoarthritis pain. *Journal of Clinical Pharmacology, 44*, 368-372.

Bargagliotti, T., Luttrell, M. F., & Lenburg, C. B. (1999). *Reducing threats to the implementation of a Competency-Based Performance Assessment System* [Electronic Journal]. Online Journal of Issues in Nursing. Retrieved 2 November, 2001, from the World Wide Web: http://www.nursingworld.or/ojin/topic10/tpc10_5.htm

Barnard, J. J. (1999). Item analysis in test construction. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 195 - 206). Oxford: Elsevier Science Pty Ltd.

Barnhardt, R. (no date). *Behaviourally Anchored Rating Scales* [web page]. COMPET Consulting. Retrieved September 18, 2002, from the World Wide Web: http://www.competinc.com/article2.html

Barrows, H. S., Williams, R. G., & Moy, R. H. (1987). A comprehensive performance-based assessment of fourth-year students' clinical skills. *Journal of Medical Education, 62*, 805 - 809.

Beeston, S., & Higgs, J. (2001). Professional practice: Artistry and connoisseurship. In J. Higgs & A. Titchen (Eds.), *Practice Knowledge and Expertise in the Health Professions. :* (pp. 108 - 121). Oxford: Butterworth-Heinemann.

Begg, T. L., & Ferguson, A. (2004, 29 August - 2 September). *Student Clinical Performance: Relationships Between Methods of Assessment.* Paper presented at the International Association of Logopedics and Phoniatrics World Congress, Brisbane.

Benner, P. (1984). *From Novice to Expert: Excellence and power in clinical nursing practice* (Commemorative issue 2001 ed.). New Jersey: Prentice-Hall.

Benner, P. A., Tanner, C. A., & Chesla, C. A. (1996). *Expertise in Nursing Practice: Caring, Clinical Judgment, and Ethics.* New York: Springer Publishing Company.

Best, D., & Rose, M. (1996). *Quality Supervision: Theory and practice for clinical supervisors.* London: WB Saunders.

Biggs, J. B., & Collis, K. F. (1982). *Evaluating the Quality of Learning: The SOLO taxonomy (Structure of the Observed Learning Outcome)*. New York: Academic Press.

Bitzer, E. (1999). *Assessing Learning in the MPhil (Higher Education Studies): Paving new ways with cornerstones.* Paper presented at the Higher Education Research and Development Society of Australasia Annual International Conference, Melbourne.

Bloom, B. S. (1994). Reflections on the development and use of the taxonomy. In L. W. Anderson & L. A. Sosniak (Eds.), *Bloom's Taxonomy: A forty-year retrospective* (pp. 1-8). Chicago: The University of Chicago Press.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*: Mahwah, NJ: Lawrence Erlbaum and Associates.

Boshuizen, H. P. A., & Schmidt, H. G. (2000). The development of clinical reasoning expertise. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed., pp. 15 - 22). Oxford: Butterworth-Heinemann.

Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education, 22*(2), 151-167.

Boyatzis, R. E., & Goleman, D. (2001). The Emotional Competence Inventory - University Edition (pp. 24). Boston: Hay Group.

Brasseur, J. (1989). The supervisory process: A continuum perspective. *Language, Speech and Hearing Services in Schools, 20*, 274-295.

Brookhart, S. M. (2001). Successful Students' Formative and Summative Uses of Assessment Information. *Assessment in Education, 8*(2), 153 - 168.

Brualdi, A. (1999, 12/99). *Traditional and Modern Concepts of Validity.* [Web document]. ERIC Clearinghouse on Assessment and Evaluation, Washington DC. Retrieved 13/02/02, 2002, from the World Wide Web: http://www.ericdigests.org/2000-3/validity.htm

Bruner, J. (1983). *Child's Talk: Learning to use language*. New York: W. W. Norton & Company.

Candy, P., & Worral-Carter, L. (1999). Educating health science students for lifelong learning. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners* (pp. 159-165). Oxford: Butterworth-Heinemann.

Carraccio, C., Wolfsthal, S. D., Englander, R., Ferentz, K., & Martin, C. (2002). Shifting paradigms: From Flexner to competencies. *Academic Medicine, 77*(5), 361-367.

Carter, R. (1985). A taxonomy of objectives for professional education. *Studies in Higher Education, 10*(2), 135-149.

Chapman, J. (1998). Agonising about assessment. In D. Fish & C. Coles (Eds.), *Developing Professional Judgement in Health Care: Learning through the critical appreciation of practice* (pp. 157-181). Oxford: Butterworth-Heinemann.

Clark, D. (1999, 21/05/00). *Learning Domains or Bloom's Taxonomy* [web page]. Retrieved 20/06/02, 2002, from the World Wide Web: http://www.nwlink.com/~donclark/hrd/bloom.html

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment, 7*(3), 309-319.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24*(4), 310.

Cohen, R., Rothman, A. I., Poldre, P., & Ross, J. (1991). Validity and generalizability of global ratings in an objective structured clinical examination. *Academic Medicine, 66*(9), 545-548.

Cook, K. F., Ashton, C. M., Byrne, M. M., Brody, B., Geraci, J., Giesler, R. B., Hanita, M., Souchek, J., & Wray, N. P. (2001). A psychometric analysis of the measurement level of the rating scale, time trade-off, and standard gamble. *Social Science and Medicine, 53*, 1275-1285.

Cox, K. (2000). Examining and recording clinical performance: A critique and some recommendations. *Education for Health, 13*(1), 45-52.

Cross, V. (1998). Begging to differ? Clinicians' and academics' views on desirable attributes for physiotherapy students on clinical placement. *Assessment & Evaluation in Higher Education, 23*(3), 295-311.

Cross, V., Hicks, C., & Barwell, F. (2001). Exploring the gap between evidence and judgement: Using video vignettes for practice-based assessment of physiotherapy undergraduates. *Assessment & Evaluation in Higher Education, 26*(3), 189-212.

Crossley, J., Humphris, G., & Jolly, B. C. (2002). Assessing health professionals. *Medical Education, 36*, 800-804.

Crotty, M. (1998). *The Foundations of Social Research: Meaning and perspective in the research process*. St Leonards: Allen and Unwen Pty Ltd.

Curtis, D. D. (2004). Person misfit in attitude surveys: Influences, impacts and implications. *International Education Journal, 5*(2), 125-143.

Curtis, D. D., & Boman, P. (2004). *The Identification of Misfitting Response Patterns to, and Their Influences on the Calibration of, Attitude Survey Instruments.* Paper presented at the 13th International Objective Measurement Workshop, Cairns, QLD.

Curtis, D. D., & Denton, R. (2002). *The Authentic Performance-based Assessment of Problem Solving (Draft)*: National Council for Vocational Education Research Ltd.

Dauphinee, W. D. (1995). Assessing clinical performance: Where do we stand and what might we expect? *The Journal of the American Medical Association, 274*(9), 741-743.

Davis, J. (2002). Comparison of faculty, peer, self, and nurse assessment of obstetrics and gynecology residents. *Obstetrics and Gynecology, 99*(4), 647-651.

Dawson, V. (1993). Competency based standards for speech pathologists. *Australian Communication Quarterly*(Autumn), 9-10.

de Laine, M. (1997). *Ethnography: Theory and applications in health research*. Sydney: MacLennan & Petty Pty Ltd.

Delbridge, A., Bernard, J. R. L., Bauer, L., Butler, S., Hodges, F., Atkinson, A., Lambert, J., & Moore, A. (1981). *The Macquarie Dictionary*. Sydney: Macquarie Library Pty Ltd.

Down, C., & Hager, P. (1999). *Making Judgements: Practical strategies from research outcomes.* Paper presented at the Australian Vocational Education and Training Research Association Conference, Melbourne.

Down, C., Martin, E., Hager, P., & Bricknell, L. (1999). *Graduate Attributes, Key Competence and Judgements: Exploring the links.* Paper presented at the Higher Education Research and Development Society of Australasia Annual International Conference, Melbourne.

Dreyfus, H. L., & Dreyfus, S. E. (1996). The relationship of theory and practice in the acquisition of skill. In P. A. Benner & C. A. Tanner & C. A. Chesla (Eds.), *Expertise in Nursing Practice: Caring, clinical judgment, and ethics.* (pp. 29-47). New York: Springer Publishing Company.

Duke, M. (1996). Clinical evaluation - difficulties experienced by sessional clinical teachers of nursing: A qualitative study. *Journal of Advanced Nursing, 23*(2), 408-414.

Edelstein, R. A., Reid, H. M., Usatine, R., & Wilkes, M. S. (2000). A comparative study of measures to evaluate medical students' performances. *Academic Medicine, 75*(8), 825-833.

Education, Q. A. A. f. H. (2001). *Academic and Practitioner Standards: Speech and Language Therapy*. Gloucester: Quality Assurance Agency for Higher Education.

Embretson (Whitely), S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197.

Embretson, S. E. (1999). Issues in the measurement of cognitive abilities. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement: What every psychologist and educator should know* (pp. 1-16). Mahwah: Lawrence Erlbaum Associates, Inc.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists.* Mahwah: Lawrence Erlbaum Associates.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *Journal of the American Medical Association, 287*(2), 226-235.

Eraut, M. (1994). *Developing Professional Knowledge and Competence.* London: The Falmer Press.

Eraut, M. (1998). Concepts of competence. *Journal of Interprofessional Care, 12*(2), 127-139.

Fay, C. H., & Latham, G. P. (1982). Effects of training and rating scales on rating errors. *Personnel Psychology, 35,* 105-117.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (pp. 105-145). New York: Macmillan Publishing Co.

Ferguson, A., & Elliot, N. (2001). Analysing aphasia treatment sessions. *Clinical Linguistics & Phonetics, 15*(3), 229 - 243.

Ferguson, A., & Fitzpatrick-Barr, K. (2001). *Awareness of Readiness for Self-directed Learning: A pilot study.* Paper presented at the Speech Pathology Australia National Conference, Melbourne.

Ferguson, A., Gibbons, J., Van Der Wal, A., James, C., & Baines, S. (2001). *Critical Thinking: Processes and Outcomes in Education.* Paper presented at the Speech Pathology Australia National Conference, Melbourne.

Fish, D., & Coles, C. (1998). Giving professionalism back to professionals. In D. Fish & C. Coles (Eds.), *Developing Professional Judgement in Health Care: Learning through the critical appreciation of practice* (pp. 289-307). Oxford: Butterworth-Heinemann.

Fisher, J. (1998). Assessment of clinical competency in sonography in the United Kingdom. *Journal of Diagnostic Medical Sonography, 14*(4), 169-171.

Fisher, J., W P. (1994). The Rasch debate: Validity and revolution in educational measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 36-72). Norwood: Ablex Publishing Corporation.

Fleming, M. H., & Mattingly, C. (2000). Action and narrative: Two dynamics of clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed.). Oxford: Butterworth-Heinemann.

Fontaine, S., & Wilkinson, T. J. (2003). Monitoring medical students' professional attributes: Development of an instrument and process. *Advances in Health Sciences Education, 8,* 127-137.

French, S., Reynolds, F., & Swain, J. (2001). *Practical Research: A guide for therapists.* Oxford: Butterworth-Heinemann.

Friedman, M., & Mennin, S. (1991). Rethinking critical issues in performance assessment. *Academic Medicine, 66*(7), 390 - 395.

Gamble, J., Chan, P., & Davey, H. (2001). Reflection as a tool for developing professional practice knowledge and expertise. In J. Higgs & A. Titchen (Eds.), *Practice Knowledge and Expertise in the Health Professions.* (pp. 121 -128). Oxford: Butterworth-Heinemann.

Geffen, L. (1992). Viewpoint. *University News*, 6.

Gomez-Mejia, G. (1988). Evaluating employee performance: Does the appraisal instrument make a difference? *Journal of Organizational Behavior Management, 9*(2), 155-172.

Gonczi, A. (1992). *A Guide to the Development of Competency Standards for the Professions*. Canberra: National Office for Overseas Skills Recognition.

Govaerts, J. J. B., van der Vleuten, C., & Schuwirth, L. W. T. (2002). Optimising the reproducibility of a performance-based test in midwifery education. *Advances in Health Sciences Education, 7,* 133-145.

Grant, J. (1999). The incapacitating effects of competence: A critique. *Advances in Health Sciences Education, 4*(3), 271-277.

Gronlund, N. E. (2003). *Assessment of Student Achievement* (7th ed.). Boston: Allyn & Bacon.

Hager, P. (1999). *Making Judgments as the Basis for Workplace Learning - Preliminary Research Findings.* Paper presented at the Australian Vocational Education and Training Research Association Conference, Melbourne.

Hager, P. (2000). Know-how and workplace practical judgement. *Journal of Philosophy of Education, 34*(2), 281-296.

Hager, P., Athanasou, J., & Gonczi, A. (1994). *Assessment Technical Manual*. Canberra: Australian Government Publishing Service.

Harris, I. (1993). New expectations for professional competence. In L. Curry & J. F. Wergin (Eds.), *Educating Professionals* (pp. 17-52). San Francisco: Jossey-Bass.

Harris, R., Guthrie, H., Hobart, B., & Lundberg, D. (1995). *Competency-based Education and Training: Between a rock and a whirlpool*. Sydney: Macmillan Education Australia Pty Ltd.

Hayes, K. W., Huber, G., Rogers, J., & Sanders, B. (1999). Behaviors that cause clinical instructors to question the clinical competence of physical therapist students. *Physical Therapy, 79*(7), 653-667.

Hays, R. B., Davies, H. A., Beard, J. D., Caldon, L. F. M., Farmer, E. A., Finucane, P. M., McCrorie, P., Newble, D., Schuwirth, L. W. T., & Sibbald, G. R. (2002). Selecting performance assessment methods for experienced physicians. *Medical Education, 36*, 910-917.

Hays, R. B., Jolly, B. C., Caldon, L. F. M., McCrorie, P., McAvoy, P. A., McManus, I. C., & Rethans, J.-J. (2002). Is insight important? Measuring capacity to change performance. *Medical Education, 36*, 965-971.

Henley, E., & Twible, R. (2000). Teaching clinical reasoning across cultures. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed., pp. 255-261). Oxford: Butterworth-Heinemann.

Higgs, J. (1997). Learning to make clinical decisions. In L. McAllister & M. Lincoln & S. McLeod & D. Maloney (Eds.), *Facilitating Learning in Clinical Settings* (pp. 130-153). Cheltenham: Stanley Thomas Ltd.

Higgs, J. (1999, 27-29 September). *Doing, Knowing, Being and Becoming in Professional Practice.* Paper presented at the MTeach Post Internship Conference, The University of Sydney, Sydney University.

Higgs, J., & Bithell, C. (2001). Professional expertise. In J. Higgs & A. Titchen (Eds.), *Practice Knowledge and Expertise in the Health Professions.* (pp. 59-68). Oxford: Butterworth-Heinemann.

Higgs, J., & Edwards, H. (1999). Educating beginning practitioners in the health professions. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners* (pp. 3 - 9). Oxford: Butterworth-Heinemann.

Higgs, J., & Hunt, A. (1999). Rethinking the beginning practitioner: introducing the 'Interactional Professional". In J. Higgs & A. Hunt (Eds.), *Educating Beginning Practitioners* (pp. 21-29). Oxford: Butterworth-Heinemann.

Higgs, J., & Jones, M. (2000). Clinical reasoning in the health professions. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed., pp. 3 - 14). Oxford: Butterworth-Heinemann.

Higgs, J., Jones, M., & Refshauge, K. (1999). Helping students learn clinical reasoning skills. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners* (pp. 197 - 203). Oxford: Butterworth-Heinemann.

Higgs, J., & Titchen, A. (2001). *Practice Knowledge and Expertise in the Health Professions.* Oxford: Butterworth-Heinemann.

Higgs, J., Titchen, A., & Neville, V. (2001). Professional practice and knowledge. In J. Higgs & A. Titchen (Eds.), *Practice Knowledge and Expertise in the Health Professions.* (pp. 3 - 9). Oxford: Butterworth-Heinemann.

Hrachovy, J., Clopton, N., Baggett, K., Garber, T., Cantwell, J., & Schreiber, J. (2000). Use of the Blue MACS: Acceptance by clinical instructors and self-reports of adherence. *Physical Therapy, 80*(7), 652-661.

Hunt, A., Adamson, B., & Harris, L. (1999). Community and workplace expectations of health science graduates. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners* (pp. 38-45). Oxford: Butterworth-Heinemann.

Hunt, A., & Higgs, J. (1999). Learning generic skills. In J. Higgs & H. Edwards (Eds.), *Educating Beginning Practitioners* (pp. 166-172). Oxford: Butterworth-Heinemann.

Hunt, D. D. (1992). Functional and dysfunctional characteristics of the prevailing model of clinical evaluation systems in North American medical schools. *Academic Medicine, 67*(4), 254-259.

Ilott, I., & Murphy, R. (1997). Feelings and failing in professional training: The assessor's dilemma. *Assessment and Evaluation in Higher Education, 22*(3), 307 - 316.

Johnson, C. J., & Shewan, C. M. (1988). A new perspective in evaluating clinical effectiveness: The UWO clinical grading system. *Journal of Speech and Hearing Disorders, 53*, 328-340.

Johnson, J. M. (1997). *Visual Analog Scales: Part I* [web page]. Department of Clinical Investigation, Brook Army Medical Centre. Retrieved September 18, 2002, from the World Wide Web: http://www.bamc.amedd.army.mil/DCI/articles/dci04974.htm

Jones, A. (2000). The place of judgement in competency-based assessment. *Journal of Vocational Education and Training, 51*(1), 145-160.

Jones, A. (2001a, July 10th to 13th). *I don't care just as long as it looks yellow.* Paper presented at the Australian Vocational Education and Training Research Association Conference, Adelaide.

Jones, A. (2001b). *It's a judgement call ... and consistency isn't all it's cracked up to be.* Paper presented at the Australian Vocational Education and Training Research Association Conference, Adelaide.

Kane, M. T. (1992). The assessment of professional competence. *Evaluation and the Health Professions, 15*(2), 163-182.

Keen, A. J. A., Klein, S., & Alexander, D. A. (2003). Assessing the communication skills of doctors in training: Reliability and sources of error. *Advances in Health Sciences Education, 8*, 5-16.

Kingstrom, P. O., & Bass, A. R. (1981). A critical analysis of studies comparing Behaviorally Anchored Rating Scales (BARS) and other rating formats. *Personnel Psychology, 34*(2), 263-289.

Krathwohl, D. R. (1994). Reflections on the taxonomy: Its past, present and future. In L. W. Anderson & L. A. Sosniak (Eds.), *Bloom's Taxonomy: A forty-year retrospective* (pp. 181-202). Chicago: The University of Chicago Press.

Krueger, D. L., & Morgan, R. A. (1998). *The Focus Group Kit*. Thousand Oaks: SAGE Publications Inc.

Landy, F. J., & Farr, J. L. (1980). Performance rating. *Psychological Bulletin, 87*(1), 72-107.

Leach, L., Neutze, G., & Zepke, N. (2001). Assessment and empowerment: Some critical questions. *Assessment & Evaluation in Higher Education, 26*(4), 293-305.

Lew, S. R., Page, G. G., Schuwirth, L. W. T., Baron-Maldonado, M., Lescop, J., Paget, N., Southgate, L. J., & Wade, W. B. (2002). Procedures for establishing defensible programmes for assessing practice performance. *Medical Education, 36*, 936-941.

Linacre, J. M. (1994). *Many-Facet Rasch Measurement* (3rd Ed ed.). Chicago: MESA Press.

Linacre, J. M. (1995). Categorical misfit statistics. *Rasch Measurement Transactions [electronic journal], 2004*(16 July).

Linacre, J. M. (1998a). Rating, judges and fairness. *Rasch Measurement Transactions [electronic journal], 2004*(10 August).

Linacre, J. M. (1998b). Visual analog scales. *Rasch Measurement Transactions [electronic journal], 12*(2).

Linacre, J. M. (1999a). Category disordering vs. step (threshold) disordering. *Rasch Measurement Transactions [electronic journal], 2004*(16 July).

Linacre, J. M. (1999b). Meditations on the Handbook of Modern Item Response Theory. *Rasch Measurement Transactions [electronic journal], 2004*(7 July).

Linacre, J. M. (2001). Glossary of Rasch measurement terminology. *Rasch Measurement Transactions [electronic journal], 2004*(9 October).

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M., & Wright, B. D. (1998). Bigsteps [Statistical]. Chicago: MESA Press.

Linacre, J. M., & Wright, B. D. (2003). *A User's Guide to Bigsteps: Rasch-model computer program*: Winsteps.com.

Lincoln, M. (2002). *Learning Time Management Skills: Why? Where? When? And How?* Paper presented at the Speech Pathology Australia National Conference, Alice Springs.

Ling, P. (1999, July, 1999). *Assessing Competency.* Paper presented at the Higher Education Research and Development Society of Australasia Annual International Conference, Melbourne.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21.

Loomis, J. (1985a). Evaluating clinical competence of physical therapy students. Part 2: Assessing the reliability, validity and usability of a new instrument. *Physiotherapy Canada, 37*(2), 91-98.

Loomis, J. (1985b). Evaluating clinical competence of physical therapy students. Part 1: The development of an instrument. *Physiotherapy Canada, 37*(2), 83-89.

Lopez, W. A. (1996). Communication validity and rating scales. *Rasch Measurement Transactions [electronic journal], 2004*(16 July).

Luttrell, M. F., Lenburg, C. B., Scherubel, J. C., Jacob, S. R., & Kock, R. (1999). Competency outcomes for learning and performance assessment: Redesigning a BSN curriculum. *Nursing and Health Care Perspectives, 20*(3), 134-141.

Maclellan, E. (2001). Assessment for learning: the differing perceptions of tutors and students. *Assessment & Evaluation in Higher Education, 26*(4), 307-318.

Maloney, D., Carmody, D., & Nemeth, E. (1997). Students experiencing problems learning in clinical settings. In L. McAllister & M. Lincoln & S. McLeod & D. Maloney (Eds.),

Facilitating Learning in Clinical Settings (pp. 185-213). Cheltenham: Stanley Thornes Ltd.

Masters, G. N. (1999). Measuring performance: The challenge of assessment. *Independent Education, 29*(1), 18 - 21.

Masters, G. N., Adams, R. J., & Wilson, M. (1999). Charting of student progress. In G. N. Masters & J. P. Keeves (Eds.), *Advances in Measurement in Educational Research and Assessment* (pp. 254-267). Oxford: Elsevier Science Pty Ltd.

Mattingly, C. (1991). What is clinical reasoning? *The American Journal of Occupational Therapy, 45*(11), 979-986.

McAllister, L. (1997). An adult learning framework for clinical education. In L. McAllister & M. Lincoln & S. McLeod & D. Maloney (Eds.), *Facilitating Learning in Clinical Settings* (pp. 1-26). Cheltenham: Stanley Thorns Ltd.

McAllister, L., Barrie, S., Mortensen, L., with, i. c., Worrall, L., Robertson, C., Russell, A., McAllister, S., Franke, M., Dann, N., & Dawson, V. (1996). *Developing Professional Competency: Self-directed learning modules for speech pathology students. Module 2: Indicators of Emerging Competence for the Competency-Based Occupational Standards for Speech Pathologists - Entry Level*: Produced with funding from a National Teaching Development Grant, Committee for the Advancement of University Teaching.

McAllister, L., & Lincoln, M. (2004). Development of personal skills, *Clinical Education in Speech-Language Pathology* (pp. 102-124). London: Whurr Publishers Ltd.

McAllister, L., & Rose, M. (2000). Speech-language pathology students: Learning clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed., pp. 204-213). Oxford: Butterworth-Heinemann.

McAllister, S. M., & Brown, R. I. (1999). User evaluation of disability services: A case study from the Guide Dog Association of South Australian and Northern Territory. *International Journal of Practical Approaches to Disability, 23*(1).

McCormack, B., & Titchen, A. (2001). Patient-centred practice: an emerging focus for nursing expertise, *Practice Knowledge and Expertise in the Health Professions.* (pp. 96 - 101). Oxford: Butterworth-Heinemann.

McGaghie, W. C. (1993). Evaluating competence for professional practice. In L. Curry & J. F. Wergin (Eds.), *Educating Professionals* (pp. 229-261). San Francisco: Jossey-Bass.

McGuire, C. H. (1995). Reflections of a maverick measurement maven. *The Journal of the American Medical Association, 274*(9), 735-740.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: Macmillan Publishing Co.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.

Messick, S. (1996). Validity of performance assessments. In G. W. Phillips (Ed.), *Technical Issues in Large-Scale Performance Assessment* (pp. 1-18). Washington: National Centre for Education Statistics.

Michell, J. (1997). Quantitative science and the definition of *measurement* in psychology. *British Journal of Psychology, 88*, 355-383.

Miller, A. H., Imrie, B. W., & Cox, K. (1998). *Student Assessment in Higher Education.* London: Kogan Page Ltd.

Miller, G. (1990). The Assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63 - S67.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24*(4), 367-377.

Milton, J. (1999). *Interpreting Competence Differently: Lessons for university strategy.* Paper presented at the Higher Education Research and Development Society of Australasia Annual International Conference, Melbourne.

Morris, M., Porter, A., & Griffiths, D. (2003). *Assessment as a Tool for Learning.* Paper presented at the Evaluations and Assessment Conference, Adelaide, South Australia.

Mosenkis, J. (1997). Recoding and pivoting: An example. *Rasch Measurement Transactions [electronic journal], 2004*(5 November).

Moss, P. (1994). Can there be validity without reliability? *Educational Researcher, 23*(2), 5-12.

Munshi, J. (1990). *A Method for Constructing Likert Scales: Research report*: Sonoma State University, CA.

Muraki, E., Hombo, C. M., & Lee, Y. (2000). Equating and linking of performance assessments. *Applied Psychological Measurement, 24*(4), 325.

Neary, M. (2000a). Responsive assessment of clinical competence: Part 1. *Nursing Standard, 15*(9), 34-36.

Neary, M. (2000b). Supporting students' learning and professional development through the process of continuous assessment and mentorship. *Nurse Education Today, 20*, 463-474.

Newble, D., Norman, G., & van der Vleuten, C. (2000). Assessing clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed., pp. 156-165). Oxford: Butterworth-Heinemann.

Newble, D., & Swanson, D. B. (1988). Psychometric characteristics of the objective structured clinical examination. *Medical Education, 22,* 325-334.

Newble, D., van der Vleuten, C., & Norman, G. (1995). Assessing clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (pp. 168-178). Oxford: Butterworth-Heinemann.

NMBE. (2002). *Embedding Professionalism in Medical Education: Assessment as a tool for implementation*. Baltimore, Maryland: National Board of Medical Examiners.

Norman, G., van der Vleuten, C., & de Graaff, E. (1991). Pitfalls in the pursuit of objectivity: Issues of reliability. *Medical Education, 25,* 110-118.

Norman, I. J., Watson, R., Murrells, T., Calman, L., & Redfern, S. (2002). The validity and reliability of methods to assess the competence to practise of pre-registration nursing and midwifery students. *International Journal of Nursing Studies., 39,* 133-145.

O'Donohue, W. J., & Wergin, J. F. (1978). Evaluation of medical students during a clinical clerkship in internal medicine. *Journal of Medical Education, 53,* 55-58.

Page, G. G. (2004). *Assessment of Fitness to Practice.* Paper presented at the Australian and New Zealand Association of Medical Education Annual Conference 2004, Adelaide, South Australia.

Pearce, R. (2001). *Performance Level Assessment in TAFE Queensland: Project Overview 2000 - 2001.* Paper presented at the Up-Grading Assessment, Melbourne.

Peters, G., Fraser, J., Cowie, F., Loader, J., Rutter, S., Scott, R., & Davie, B. (2001). Competency based assessment in a perioperative nursing graduate diploma. *ACORN Journal, 15*(5), 11-14.

Pithers, B. (2000). *The Importance of Facilitating Critical Reasoning in the New Millennium: Some new evidence.* Paper presented at the 29th International Teaching and Learning Conference, Frankfurt, Germany.

Powell, C. V., Kelly, A., & Williams, A. (2001). Determining the minimum clinically significant difference in visual analog pain score for children. *Annals of Emergency Medicine, 37*(1), 28-31.

Priest, H., & Roberts, P. (1998). Assessing students' clinical performance. *Nursing Standard, 12*(48), 37-41.

QAAHE. (2001). *Academic and Practitioner Standards: Speech and Language Therapy*. Gloucester: Quality Assurance Agency for Higher Education.

Ramsey, P. G., Wenrich, M. D., Carline, J. D., Inui, T. S., Larson, E. B., & LoGerfo, J. P. (1993). Use of peer ratings to evaluate physician performance. *The Journal of the American Medical Association, 269*(13), 1655-1660.

Refshauge, K., & Higgs, J. (2000). Teaching clinical reasoning. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed., pp. 141-147). Oxford: Butterworth-Heinemann.

Rethans, J.-J., Norcini, J., Baron-Maldonado, M., Blackmore, D., Jolly, B. C., LaDuca, T., Lew, S. R., Page, G. G., & Southgate, L. J. (2002). The relationship between competence and performance: Implications for assessing practice performance. *Medical Education, 36*, 901-909.

Rheault, W., & Coulson, E. (1991). Use of the Rasch model in the development of a clinical competence scale. *Journal of Physical Therapy Education, 5*(1), 10-13.

Roach, K., Gandy, J., Deusinger, S., Clark, S., Gramet, P., Gresham, B., Hagler, P., Lewthwaite, R., May, B. J., Sanders, B., Strube, M. J., & Rainey, Y. (2002). The development and testing of APTA Clinical Performance Instruments. *Physical Therapy, 82*(4), 329-353.

Robertson, I., Simons, M., & Harris, R. (2000). *Learning and Assessment Issues in Apprenticeships and Traineeships.* Paper presented at the Australian Vocational Education and Training Research Association Conference, Canberra.

Robertson, S., Rosenthal, J., & Dawson, V. (1997). Using assessment to promote student learning, *Facilitating Learning in Clinical Settings* (pp. 154 -184). Cheltenham: Stanley Thornes Ltd.

Rubin, J. (1996). Impediments to the development of clinical knowledge and ethical judgement in critical care nursing. In P. A. Benner & C. A. Tanner & C. A. Chesla (Eds.), *Expertise in Nursing Practice: Caring, Clinical Judgment, and Ethics.* (pp. 170-192). New York: Springer Publishing Company.

Ryan, S. (2000). Facilitating the clinical reasoning of occupational therapy students on fieldwork placement. In J. Higgs & M. Jones (Eds.), *Clinical Reasoning in the Health Professions* (2nd ed., pp. 242-248). Oxford: Butterworth-Heinemann.

Schiavetti, N., & Metz, D. E. (1997). *Evaluating Research in Communicative Disorders.* Boston: Allyn and Bacon.

Scholten, I. (2000). Development of a Test of Speech Pathology Students' Knowledge of Essential Aspects of the Normal Swallowing Process: Unpublished manuscript.

Schon, D. A. (1987). *Educating the Reflective Practitioner: Toward a new design for teaching and learning in the professions*. San Francisco: Jossey-Bass Inc.

Schuwirth, L. W. T., Southgate, L. J., Page, G. G., Paget, N., Lescop, J., Lew, S. R., Wade, W. B., & Baron-Maldonado, M. (2002). When enough is enough: A conceptual basis for fair and defensible practice performance assessment. *Medical Education, 36,* 925-930.

Schuwirth, L. W. T., & van der Vleuten, C. (2003). ABC of learning and teaching in medicine: Written assessment. *British Medical Journal, 326,* 643-645.

Schwabbauer, M. (2000). But can they do it? Clinical competency assessment. *Clinical Laboratory Science, 13*(1), 47-52.

Sefton, A. J. (2001). Integrating knowledge and practice in medicine. In J. Higgs & A. Titchen (Eds.), *Practice Knowledge and Expertise in the Health Professions.* (pp. 29 - 34). Oxford: Butterworth-Heinemann.

Sharpley, B. E. (1997, 22 November 1997). *Key Competencies Project* [web document]. Monash University. Retrieved 10 May, 2003, from the World Wide Web: http://www.education.monash.edu.au/projects/kc/

Shavelson, R. J., Gao, X., & Baxter, G. (1993). *Sampling Variability in Performance Assessments: CSE Technical Report 361*. Santa Barbara: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Smith, E. V. (2001). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 2*(3), 281-311.

Smith, E. V., Wakely, M. B., De Kruif, R. E. L., & Swartz, C. W. (2003). Optimizing rating scales for self-efficacy (and other) research. *Educational and Psychological Measurement, 63*(3), 369-391.

Smith, L. (2001, 4th May). *'Grading' CBT: Findings from Queensland research.* Paper presented at the Up-Grading Assessment, Melbourne.

Smith, R. M. (1996). Polytomous mean-square fit statistics. *Rasch Measurement Transactions [electronic journal], 2004*(16 July).

Solomon, D. J., Speer, A. J., Callaway, M. R., & Ainsworth, M. A. (1996). Dimensions of clinical competence as conceptualized by medical school faculty. *Evaluation and the Health Professions, 19*(1), 68-80.

Southgate, L. J., Campbell, M., Cox, J., Jolly, B. C., McCrorie, P., & Tombleson, P. (2001). The General Medical Council's performance procedures: The development and implementation of tests of competence with examples from general practice. *Medical Education, 35*(Supplement 1), 20-28.

Southgate, L. J., Cox, J., David, T., Hatch, D., Howes, A., Johnson, N., Jolly, B. C., Macdonald, E., McAvoy, P. A., McCrorie, P., & Turner, J. (2001). The General Medical Council's performance procedures: Peer review of performance in the workplace. *Medical Education, 35*(Supplement 1), 9-19.

Southgate, L. J., Cox, J., David, T., Howes, A., Johnson, N., Jolly, B. C., Macdonald, E., McAvoy, P. A., McCrorie, P., & Turner, J. (2001). The assessment of poorly performing doctors: The development of the assessment programmes for the General Medical Council's performance procedures. *Medical Education, 35*(Supplement 1), 2-8.

SPAA. (2001). *Competency-Based Occupational Standards for Speech Pathologists (Entry Level)*. Melbourne: Speech Pathology Association of Australia Ltd.

SPAA. (2002). *Ethics Education Package*. Melbourne: Speech Pathology Association of Australia Ltd.

SPSS. (2003). SPSS 12.01 (Version 12) [Statistical]. Chicago: SPSS Inc.

Stackhouse, J., & Furnham, A. (1983). A student-centred approach to the evaluation of clinical skills. *British Journal of Disorders of Communication, 18*(3), 171-179.

Stern, D., Baily, T., & Merrit, D. (1996). *School-to-Work Policy Insights from Recent International Developments, MDS-950* (web document). Berkley: National Center for Research in Vocational Education.

Stewart, D. W., & Shamdasani, P. (1990). *Focus Groups: Theory and practice* (Vol. 20). Newbury Park: SAGE Publications, Inc.

Stone, M. H. (2003). Substantive scale construction. *Journal of Applied Measurement, 4*(3), 282-297.

Swchwabbauer, M. (2000). But can they do it? Clinical competency assessment. *Clinical Laboratory Science, 13*(1), 47-52.

Thomee, R., Grimby, G., Wright, B. D., & Linacre, J. M. (1995). Rasch analysis of visual analog scale measurements before and after treatment of patellofemoral pain syndrome in women. *Scandinavian Journal of Rehabilitation Medicine, 27*, 145-151.

Titchen, A., & Ersser, S. J. (2001). The nature of professional craft knowledge. In J. Higgs & A. Titchen (Eds.), *Practice Knowledge and Expertise in the Health Professions.* (pp. 35 - 41). Oxford: Butterworth-Heinemann.

Tracy, S., Marino, G., Richo, K., & Daly, E. (2000). The clinical achievement portfolio: An outcomes-based assessment project in nursing education. *Nurse Educator, 25*(5).

Turnbull, J., McFadyen, J., van Barneveld, C., & Norman, G. (2000). Clinical work sampling: A new approach to the problem of in-training evaluation. *Journal of General Internal Medicine, 15*(8), 556-561.

Twible, R. L., & Henley, E. C. (2001). Transferring professional craft knowledge across cultural contexts. In J. Higgs & A. Titchen (Eds.), *Practice Knowledge and Expertise in the Health Professions.* (pp. 157 - 164). Oxford: Butterworth-Heinemann.

Wass, V., van der Vleuten, C., Shatzer, J., & Jones, J. (2001). Assessment of clinical competence. *The Lancet, 357*, 945-949.

Wewers, M. E., & Lower, N. K. (1990). A critical review of visual analogue scales in the measurement of clinical phenomena. *Research in Nursing and Health, 13*, 227-236.

Whitcombe, M. E. (2002). Competency-based graduate medical education? Of course! But how should competency be assessed? *Academic Medicine, 77*(5), 359-360.

Wilson, B. (1992). Comment. *University News*, 4.

Wilson, N. (1998). *Educational Standards and the Problem of Error*.Unpublished manuscript, Tempe.

WinGuides. (2003, 2002). *Secure Password Generator* [Web based software]. Winguides.com. Retrieved January, 2003, from the World Wide Web: http://www.winguides.com/security/password.php?guide=security

Woehr, D. J., & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Wolf, A. (1995). *Competence-Based Assessment*. Buckingham: Open University Press.

Wolfe, E. W., & Gitomer, D. H. (2001). The influence of changes in assessment design on the psychometric quality of scores. *Applied Measurement in Education, 14*(1), 91-107.

Wood, L. A., & Kroger, R. O. (2000). *Doing Discourse Analysis: Methods for studying action in talk and text.* Thousand Oaks: Sage.

Woolley, A. S. (1977). The long and tortured history of clinical evaluation. *Nursing Outlook, 25*(5).

Wright, B. D. (1999). Fundamental measurement for psychology. In S. E. Embretson & S. L. Hershberger (Eds.), *The New Rules of Measurement: What every psychologist and educator should know* (pp. 65-105). Mahwah: Lawrence Erlbaum Associates, Inc.

Wright, B. D., & Linacre, J. M. (1992). Combining and splitting categories. *Rasch Measurement Transactions [electronic journal], 2004*(9 July).

Zhu, W. (1996). Should total scores from a rating scale be used directly? *Research Quarterly for Exercise and Sport, 67*(3), 363-372.