A Comparison of Task-Specific and Dimension-Specific Assessment Centres

Duncan J. R. Jackson

Members of the Supervisory Panel

Dr. Stephen G. Atkins (Chair)
Dr. Jennifer A. Stillman
Dr. Douglas Paton
Dr. Phillip E. Lowry

The real voyage of discovery consists not in seeking new landscapes, but in having new eyes
-Marcel Proust

Massey University
COLLEGE OF HUMANITIES & SOCIAL SCIENCES

School of Psychology
Private Bag 102 904,
North Shore MSC,
Auckland,
New Zealand
Telephone: 64 9 443 9799
extn 9180
Facsimile: 64 9 441 8157

To Whom It May Concern:

This is to state that, with respect to the research conducted for the Doctoral thesis entitled *"A Comparison of Task-Specific and Dimension-Specific Assessment Centres"* carried out by Duncan John Ross Jackson, the following statements are true:

i) Reference to work other than that of the candidate has been appropriately acknowledged.

ii) The research practice and ethical policies approved by Massey University have been complied with.

iii) Although the current thesis guidelines request a word limit of 100,000, the current thesis was substantially completed prior to the introduction of this limit. (It consists of approximately 117,000 words.)

Stephen G. Atkins
Supervisor Stephen G. Atkins ***Date*** 23 Sept 2003

D.J.R. Jackson
Candidate ***Date*** 23 Sept 2003

Massey University
COLLEGE OF HUMANITIES & SOCIAL SCIENCES

School of Psychology
Private Bag 102 904,
North Shore MSC,
Auckland,
New Zealand
Telephone: 64 9 443 9799
extn 9180
Facsimile: 64 9 441 8157

To Whom It May Concern:

This is to state that the research carried out for my Doctoral thesis entitled *"A Comparison of Task-Specific and Dimension-Specific Assessment Centres"* in the School of Psychology, Massey University, Albany Campus, New Zealand, is all my own work.

This is also to certify that the thesis material has not been used for any other degree.

D.J.R. Jackson
Candidate:

Date: 23 Sept 2003

Massey University
COLLEGE OF HUMANITIES & SOCIAL SCIENCES

School of Psychology
Private Bag 102 904,
North Shore MSC,
Auckland,
New Zealand
Telephone: 64 9 443 9799
extn 9180
Facsimile: 64 9 441 8157

To Whom It May Concern:

This is to state that the research conducted for the Doctoral thesis entitled *"A Comparison of Task-Specific and Dimension-Specific Assessment Centres"* was carried out by Duncan John Ross Jackson in accordance with the University's Doctoral regulations.

Stephen G. Atkins
Supervisor

Date: 23 Sept 2003

This work is dedicated to my late grandfather, Mr. Ernest F. M. Wilson, who passed from this world at 6am on Sunday the 9th of February 2003. You were such a great and noble man, and your kindness, knowledge, wisdom, and humour will be so dearly missed. I wish I could have shared the contents of my dissertation with you, as I know you would have been keenly interested. You were one of the few people, in my younger years, who tempted me into the realisation that learning could be enjoyable. You shared your extensive knowledge of astronomy with me, and stirred a fascination, which remains today. Just prior to your passing, you said to me in your profound way; "You're my best friend". You are also my best friend, my dear grandfather. May you rest well, until we meet again.

Acknowledgements

Life-consuming ventures, such as the present Ph.D., are never performed on one's own, and due consideration must be given to all those who assisted me through this journey of discovery. From the early stages, I would like to thank Dr. Jennifer A. Stillman, A/Prof. Kerry Chamberlain, and Dr. Douglas Paton for assisting me to consolidate and formulate the methodology that would be used in my research. During the design of the assessment centres I used, I am indebted to Dr. Stephen G. Atkins, Dr. Felix E. Lopez and Dr. Phillip E. Lowry for their invaluable advice. For allowing me to gain access to organisations that use assessment centre methodology, I am grateful to Mr. Andrew Hambleton, Mr. Michael Hope, Ms. Helen Gribble, Ms. Rochelle McKay, Mr. Jason Clarke, Ms. Raewyn Bennett, Sqn Ldr Wanda Morris, Sqn Ldr Paul Gallagher, Sqn Ldr Laura Gillen, and Sqn Ldr Emma Davis. On the measures used in my research, I am grateful to Dr. Richard K. Wagner, Dr. Robert J. Sternberg, Dr. Albert Bandura, and Ms. Rebecca Tovey for their generosity. For the data analysis phase, I am indebted to Dr. Jennifer A. Stillman, Dr. Stephen G. Atkins, Dr. Robert L. Brennan, Dr. Richard J. Shavelson, Dr. Noreen M. Webb, Dr. George A. Marcoulides, Dr. John Spicer, Dr. Paul Barrett, Dr. Richard Fletcher, and Dr. John Hattie for their generous advice. For the laborious task of proofing, I am particularly indebted to Dr. Jennifer A. Stillman and Dr. Stephen G. Atkins. Thank you for your kindness, patience, and conscientiousness. In the study of assessment centres, thank you to Dr. Phillip E. Lowry, Dr. Peter Herriot, Dr. Ivan T. Robertson, and Dr. William A. Gorham for not taking the status quo at face value, and having the tenacity to stand against the prevailing view. Thank you to Dr. Filip Lievens for being at the forefront of contemporary assessment centre

research. Many thanks to Dr. Nikolaos Kazantzis for updating me on acceptable practices of assessment in clinical psychology. My gratefulness is extended to the members of my immediate family, my mother Mdm Annette M. Jackson, my father Mr. Michael J. R. Jackson, my two older brothers Mr. Hamish J. R. Jackson and Mr. Alistair J. R. Jackson and my younger sister, Ms. Daisy L. Jackson. Thank you for believing in and supporting me throughout this time. Thanks to all my wonderful friends who have supported me through this journey, particularly Ms. Stella Cho, Mr. Victor Ng, Mr. Peter Johnston, and Mr. Shane Rowe. Lastly, I am forever indebted to all the members of my supervisory panel. Thank you for being both my friends and mentors. Your guidance has helped me to open my mind to the endless possibilities that could result from the study of psychology. The research contained, herein, was approved by the Massey University Albany Campus Human Ethics Committee, MUAHEC 00/047.

Abstract

Three studies were employed to further an understanding of a measurement quandary concerning assessment centres (ACs). A common theme associated with ACs is that they do not appear to measure the trait-based variables that they purport to. To compound this mystery, ACs are found to be predictive of outcome criteria; particularly criteria related to promotion. All three studies took varying perspectives on this measurement dilemma. The first study looked at particular traits that were not formally assessed in ACs, and whether these traits explained variance in overall AC ratings. No definitive evidence was found for this notion; however, tacit knowledge appeared to be associated with a small amount of variance in overall AC ratings in one of the samples under scrutiny. The second study looked at the extent to which assessors and candidates understood the models they were assessing and were being assessed under. Neither party appeared to distinguish trait-based, task-based, or other models as being more or less appropriate. While the first and second studies acknowledged some peripheral issues in the AC literature, the third study addressed the fundamental research question. Specifically, the third study investigated whether an alternative to the prevailing trait paradigm was needed. This study compared two models of assessment in a repeated measures design. One model treated the AC data as though they comprised situationally specific behavioural samples. The second model treated the data as though they were indicative of trait-based responses. Using a generalizablity study, both models demonstrated similar psychometric characteristics, although only data treated under the situationally specific model held a conceptual justification. These findings suggest that the situationally specific task-based model presents a more appropriate means by which to treat AC ratings.

A Comparison of Task-Specific and Dimension-Specific Assessment Centres

Acknowledgements i

Abstract iii

List of Tables 5

List of Figures 7

Principal Notational Conventions 8

Chapter One: Background and Hypotheses 9

Background and History of the Assessment Centre Process 9
Group Exercises 15
Individual Exercises 15
Written Exercises 16

The Trait Paradigm 17
Construct Validity and the Exercise Effect 18
The Importance of Construct Validation in ACs 21
Construct Validation of ACs through the Nomological Network 24
Factors that May Improve AC Construct Validity: The Limited Information-Processing Model 25
Rating Dimensions Subsequent to Agreeing Upon Dimensional Ratings 27
Having Assessors Rate a Singular Dimension Across Exercises 29
Reducing Cognitive Load On Assessors and Organising Ratings 30
The Use of Video Recordings 35
Dimensional Transparency 35
Exercise Transparency and Opportunities To Express Behaviour 37
Form and Content of AC exercises 38
Factors that May Improve AC Construct Validity: The Expert Assessor Model 39
Frame of Reference Training 40
Employing Psychologists as Assessors 44
Attributing Variance to both Exercise and Dimensional Features 51
Overall Assessment Rating Integration Discussions in ACs 52
The Measurement of Latent Constructs in ACs 53
The Actual Criterion Contamination Explanation 54
The Subtle Criterion Contamination Explanation 56
The Self-Fulfilling Prophecy/Self-Efficacy Explanation 58
The Managerial Intelligence Explanation 61
The Impression Management Skill Explanation 65
Intelligence, Personality, and their Relationships with Overall Assessment Ratings (OARs) 70

The Behavioural and Interactionist Paradigms 72
The Performance Consistency Explanation 75
Evidence in Favour of a Task-Specific Approach 82

Summary 87

Overall Research Aim 88

Hypotheses 90
Study One, Hypothesis One 90
Study Two, Additional Research Question 90
Study Three, Hypothesis Two 90

Chapter Two: Study One, Latent Trait Measurement in ACs 92

Method 92
Prelude to Studies One and Two 92
Military Sample 93
Participants 93
Assessors 93
The RNZAF Selection Board 95
Selection Board Dimensions 95
Selection Board Exercises 97
Measures 99
Procedure 106
Organisational Sample 106
Participants 106
Assessors 108
The AC 108
AC Dimensions 109
AC Exercises 112

Results 113
Military Sample 114
Set One 115
Set Two 121
Supplementary Analysis for Set Two of the Military Sample 126
Organisational Sample 131

Discussion 137
Military Sample 138
Organisational Sample 139
Considerations 140
Analytical Limitations 141
Theoretical Implications 142

Chapter Three: Study Two, Perceptions of Assessors and Candidates with respect to Measurement Models 145

Method 145
Candidates 145
Assessors 145
Measures: Candidates 147
Measures: Assessors 148

Results **150**
Candidates 150
Assessors 153

Discussion **157**
Candidates 158
Assessors 158
Considerations 160
Theoretical Implications 161

Chapter Four: Study Three, A Comparison of Task-Specific and Dimension-Specific ACs **163**

Method **163**
Participants 163
Assessors 163
The AC 165
Task Analysis 166
TTA (Threshold Traits Analysis) 170
TTA Respondents 171
Summarising/Scoring Responses to the TTA 172
Presentation to the Managerial Level SME Panel 174
Classification and Extrapolation of Tasks into Dimensions 175
AC Task Ratings and Dimensions 176
AC Exercises 177
Evaluation Approach 178
Assessor Training and the Assessment Procedure 178
Procedure 181

Results **182**
Generalizability Study 184
Factor Analysis 192
Varimax Rotation 193
Direct Oblimin Rotation 196
Confirmatory Factor Analysis 198

Discussion **207**
Generalizability Study 207
Factor Analysis 212
Confirmatory Factor Analysis 214
Considerations 215
Theoretical Implications 218

Chapter Five: General Discussion **222**

References **234**

Appendix I: Pilot for Study Three **251**

Method 251

Results and Discussion 262

Appendix II: Introduction to Generalizability Theory 273

Appendix III: Abridged Assessment Centre Manual and Training Guide For Farmers Merchandiser, General Sales and One On One Sales Roles: Including General Sales Exercises 300

List of Tables

Table Number	Title	Page Number
1	*Average Convergent and Discriminant Validity Coefficients of Assessor Ratings in a Sample of AC Studies*	22
2	*Predictive Validity of Various Designs of AC Research*	56
3	*Various Criteria Used and their Predictive Validity with AC Outcomes*	58
4	*Demographic Statistics, Candidates, Study One Military Sample*	94
5	*Demographic Statistics, Assessors, Study One Military Sample*	95
6	*Demographic Statistics, Candidates, Study One Organisational Sample*	107
7	*Overall Means and Standard Deviations for Measures Employed in Set One of theMilitary Sample*	116
8	*Bivariate Correlations Between Measures Employed in Set One of the Military Sample*	117
9	*Multiple Regression Analysis for the Prediction of OARs in Set One of the Military Sample*	118
10	*Spearman's Rho Between Measures Employed in Set One of the Military Sample*	120
11	*Overall Means and Standard Deviations for Measures Employed in Set Two of the Military sample*	121
12	*Bivariate Correlations Between Measures Employed in Set Two of the Military Sample*	122
13	*Multiple Regression Analysis for the Prediction of OARs in Set Two of the Military Sample*	124
14	*Spearman's Rho Between Measures Employed in Set Two of the Military Sample*	125
15	*Bivariate Correlations Between Measures Employed in Supplementary Set Two of the Military Sample*	128
16	*Multiple Regression Analysis for the Prediction of OARs in Supplementary Set Two of the Military Sample*	129
17	*Spearman's Rho Between Measures Employed in Supplementary Set Two of the Military Sample*	130
18	*Overall Means and Standard Deviations for Measures Employed in the Organisational Sample*	133
19	*Bivariate Correlations Between Measures Employed in the Organisational Sample*	133
20	*Multiple Regression Analysis for the Prediction of OARs in the Organisational Sample*	135
21	*Spearman's Rho Between Measures Employed in the Organisational Sample*	136
22	*Demographic Statistics, Candidates, Study Two Military Sample*	146
23	*Demographic Statistics, Assessors, Study Two Military Sample*	147

24	*Model Assumed to Underlie Assessment: Candidates*	151
25	*Model Assumed to Guide Assessment: Candidates*	152
26	*Model Assumed to Guide Assessment After Being Informed of the Measurement Problems in ACs: Candidates*	153
27	*Usefulness of Individual Dimensions: Assessors*	154
28	*Dimensions Perceived as Being Seen Exhibited Across All Exercises: Assessors*	155
29	*Model Assumed to Guide Assessment: Assessors*	156
30	*Model Assumed to Guide Assessment After Being Informed Of The Measurement Problems In ACs: Assessors*	157
31	*Demographic Statistics, Candidates, Study Three Private Sector Sample*	164
32	*Grand Means and SDs of the Behavioural Ratings (Within Exercises) in the Task-Specific AC*	183
33	*Grand Means and SDs of the Dimension Ratings (Across Exercises) in the Dimension Specific AC*	183
34	*Generalizability Study Comparing a Task-Specific with a Dimension-Specific AC in a Repeated Measures Design for the Organisational Sample*	187
35	*Relative and Absolute Error, Generalizability and Phi Coefficients and Interrater Reliability for the Balanced Task-Specific AC*	190
36	*Relative and Absolute Error, Generalizability and Phi Coefficients and Interrater Reliability for the Dimension-Specific AC*	191
37	*Generalizability Study Showing the Results of the Unbalanced Task-Specific AC for the Organisational Sample*	192
38	*Rotated Factor Matrix for the Task-Specific AC Ratings*	194
39	*Rotated Factor Matrix for the Dimension-Specific AC Ratings*	195
40	*Rotated Pattern Matrix for the Task-Specific AC Ratings*	197
41	*Rotated Pattern Matrix for the Dimension-Specific AC Ratings*	198
42	*Standardised Factor Loadings for Model One: The Abridged Task-Specific CFA Model*	201
43	*Selected Goodness-Of-Fit Indices for Model One: The Abridged Task-Specific CFA Model*	202
44	*Standardised Factor Loadings for Model Two: The Dimension-Specific CFA Model*	203
45	*Selected Goodness-Of-Fit Indices for Model Two: The Dimension-Specific CFA Model*	204
46	*Standardised Factor Loadings for Model Three: The Exercise Effect CFA Model*	205
47	*Selected Goodness-Of-Fit Indices for Model Three: The Exercise Effect CFA Model*	206
48	*Advantages of the Task-Specific Approach Relative to the Dimension-Specific Approach to AC Design*	228

List of Figures

Figure Number	Caption	Page Number
1	*Competency/Exercise Matrix for Study One, Organisational Sample*	110
2	*Variance Components and Confidence Intervals for Each Effect and Interaction in the Task-Specific AC*	189
3	*Variance Components and Confidence Intervals for Each Effect and Interaction in the Dimension-Specific AC*	189
4	*Model One: Abridged Task-Specific CFA Model*	199
5	*Model Two: Dimension-Specific CFA Model*	202
6	*Model Three: The Exercise Effect CFA Model*	204

Principal Notational Conventions for Generalizability Studies

p	The main effect for persons, the object of measurement in G studies.
x	The main effect for assessment centre exercises. This and any other source of variance in a G study, except for the object of measurement, is termed a *facet*.
d	The main effect for dimensions, traits, or competencies. These constructs are assumed to have a quality that is relatively stable and enduring across assessment exercises. This and any other source of variance in a G study, except for the object of measurement, is termed a facet.
px	The interaction term for two (or more) facets in a G study.
pxd,e	The interaction between all the facets in a G study followed by an 'e' indicates the error term for the model. This is the component of variance that is attributable to undifferentiated error.
$i:x$	The presence of a colon (:) indicates that one facet is nested within another. In this case, the facet '*i*' (items) is nested within 'x' (exercises). This occurs in a task-specific assessment centre, because each exercise has its own associated set of items.
σ^2_{Rel}	Relative error term. Used to calculate measurement error associated with all of the components of variance that compare the standing of individuals relative to one another. This term is used in the calculation of the G coefficient.
σ^2_{Abs}	Absolute error term. Used to calculate measurement error associated with all of the components of variance that relate to absolute decisions. That is, decisions that have a cut-off point, or a pass/fail criterion. This term is used in the calculation of the Phi coefficient.
$E\rho^2_{Rel}$	The G coefficient for relative decisions. This is presented on a scale from 0, indicating poor generalizability, to 1, indicating excellent generalizability.
ϕ	The Phi coefficient for absolute decisions. This is presented on a scale from 0, indicating poor generalizability, to 1, indicating excellent generalizability.

Chapter One: Background and Hypotheses

Background and History of The Assessment Centre Process

Assessment centres (ACs) are a popular process used for the assessment of people in organisations. Despite the popularity of this process, ACs are frequently criticised as being anomalous with regard to the psychological constructs that they purport to measure (Chan, 1996). The history of the AC concept proper dates back to around the 1930s. After the First World War, the German Military decided to implement new selection procedures for its officers to foster a more powerful armed force for the inexorable Second World War (Seegers, 1997). The resulting selection procedure employed during the Nazi regime was the foundation for what is now known as the AC process, which involves the utilisation of multiple forms of assessment by which to acquire a standardised evaluation of an individual's behaviour (Baron & Janman, 1996). In ACs, candidates participate on several different work-related simulation exercises, which aim to measure a set of competencies relevant to a target job. Each candidate is evaluated by a number of assessors, and any given AC may last from a half day, to three days or longer. Seegers (1997) states that ACs have been used in the contexts of simulations, role-plays, selection, assessment, training, evaluation, and as a means of dealing with personnel issues.

The earliest origins of multiple assessment methodologies per se can be traced to the Han Dynasty of ancient China (BC 206 to AD 220), which employed the use of multiple testing across a diverse range of occupations. By the Ming Dynasty (AD 1368 to AD 1644), such testing procedures had become even more sophisticated, and involved multiple stages which needed to be completed consecutively before

candidates were eligible to apply for certain positions. Kaplan and Saccuzzo (2001) suggested that the Western world became familiarised with the Chinese methodologies though the verbal accounts of missionaries.

Specifically concerning AC methodology, Feltham (1989) reviewed the historical progression of the AC into the organisational context. Ironically, it was the original German selection procedures that were instrumental in the development of the next stage in the evolution of the AC: the British War Office Selection Board. The German method of selection via multiple assessment devices became widespread throughout the British military by around 1941. The British War Office Selection Board was the direct precursor to the Regular Commissions Board (in Britain) and the US Office of Strategic Services selection system, both of which benefited from moderately improved selection precision with the use of multiple assessment methodologies.

After the war came the development of the British Civil Service Selection Board, which adopted the British War Office Selection Board methodology for civil service appointments. The emphasis in this latest version of multiple assessment methodology was on the intellectual capacity of candidates. The design of the selection procedure was based on a job analysis of the civil servant's occupational characteristics. The assessment battery included cognitive ability and personality tests, interest questionnaires, projective tests, peer assessments, interviews and situational tests which were designed as simulations of the civil servant's job. All data were aggregated and an overall grade was awarded to each candidate accordingly. The British Civil Service Selection Board found the first evidence for the predictive validity of multiple assessment methodology, with training, job performance, and promotion-related criteria.

Beginning around 1948, the first adoption of multiple assessment had only a slight (although promising) realisation in selection procedures in a Scottish organisational context (Handyside & Duncan, 1954). The adopted process was very much influenced by the military, reinforced by the fact that the assessment process was constructed "with the advice and assistance of Brigadier F. I. De la P. Garforth" (Handyside & Duncan, 1954, p. 9). It is interesting to note that even at this early stage, the intention in the AC process was to measure trait-based variables such as "acceptability to others, co-operativeness, intelligence, persuasiveness and leadership" (Handyside & Duncan, 1954, p. 19).

The espousal of ACs into the organisational context was pioneered most influentially by the US company American Telephone and Telegraph (AT&T) in 1956 (Seegers, 1997). AT&T included the multiple assessment procedure in their longitudinal research into the development of employees. In the original AT&T AC, simulations, group assignments, interviews and tests were used to predict the promotion of candidates to a middle management position. The procedure consisted of a three and a half day AC, which included a two-hour interview, an "in-basket" work sample exercise, a group simulation (which referred to a manufacturing problem), a group discussion (which focussed on discussions surrounding a promotion decision), psychological tests, projective tests and questionnaires (Feltham, 1989). The AT&T procedure, which was used only for research purposes, yielded impressive results. The AC was reported as correctly predicting the promotion of 42% of candidates into middle management positions (Bray & Grant, 1966). The AT&T template for the AC procedure was followed by thirteen organisations by 1969 (Seegers, 1997).

The findings pertaining to the predictive validity of the AC process are mixed. Mean corrected predictive validity coefficients for the criterion of job performance from three meta-analytic studies have been reported as .41 (Schmitt, Gooding, Noe & Kirsch, 1984) and .37 (Gaugler, Rosenthal, Thornton & Bentson, 1987). The median corrected correlation rose to .63 for promotion data and .33 (.43 when corrected for attenuation) for performance across 21 studies (Cohen, Moses & Byham, 1974).

In a study of the ratings of school administrators using the same AC across different locations, the corrected job performance validity coefficient dropped to .16 with teacher ratings, and the uncorrected average coefficient for the ratings from the school administrator's immediate supervisors was .08 (Schmitt, Schneider & Cohen, 1990). Ostensibly, these coefficients are grounds for concern. However, it should be noted that in the Schmitt et al. study, there was an enormous amount of variability across the coefficients obtained for teacher ratings (minimum = -.40, maximum = .82) and for supervisory ratings (minimum = -.50, maximum = .64). Note also that the supervisory ratings were not corrected for unreliability associated with criterion measures. This study may also highlight some problems associated with attempting to use the same AC across different groups of assessees. Ideally, ACs should be constructed using standardised methodology, however, they should also be tailored to the requirements of specific organisations and based on organisation-specific job analyses (Joiner, 2002).

Fleenor (1996) makes the point that since a lack of standardisation across ACs exists; meta-analysis may not be the best method by which to assess the predictive validity of the process. The strength of a predictive validity coefficient also depends on the type of criterion used (Muchinsky, 2000). It should also be noted with caution that many procedures labelled "ACs" are, in practice, poorly designed and have not

concentrated on the fundamental theoretical foundations of the procedure, such as carrying out job analyses, designing exercises in accordance with current guidelines, and the training of assessors (Feltham, 1989; Fletcher & Anderson, 1998; International Task Force on Assessment Center Guidelines, 2000; Lievens, 2002). The lack of high quality job analyses for justification and support is unfortunately common for many selection systems, and this may be especially problematic in New Zealand where selection systems lack structured job analyses (Taylor, Keelty & McDonnell, 2002). In saying this, however, most of the New Zealand based respondents to Taylor et al.'s survey reported using either a job analysis or an "understanding of competencies or tasks relating to target jobs" (p. 14) specifically to design AC procedures.

The contemporary processes, now commonly called assessment and development centres, focus on standardised evaluations of behaviour, based on multiple samples of behaviour, and using multiple traits in order to assess that behaviour (Ballantyne & Povah, 1995). Development centres focus more on effective developmental responses, as opposed to being purely diagnostic (Carrick & Williams, 1999). To encapsulate both assessment and development centres, the more general term AC will be used in this review, as both use identical design elements and fundamental methods, and only differ substantively in regard to their overall aims. Ballantyne and Povah (1995) outlined the major components of ACs as involving the participation of a group in multiple exercises, who are observed and rated against predetermined task-related behaviours by a team of trained assessors. The data obtained by the assessors are then shared before final decisions are made. ACs are based on the inclusion of multiple participants, and a combination of methods by which to assess behaviour so that a complete profile can be compiled. They utilise

multiple assessors and ideally, the entire assessment process should form its basis from a job analysis of the target job or task. ACs can also include supplementary measures such as any combination of cognitive ability tests, interviews and personality questionnaires, in addition to simulation exercises, as part of the entire assessment process (Cook, 1998). This review will focus on the measurement of human attributes through simulation exercises, which represent the quintessence of AC methodology (International Task Force on Assessment Center Guidelines, 2000).

Baron and Janman (1996), in their review of fairness in the AC, report that primarily, ACs comprise sets of work sample type exercises including the in-tray exercise, the leaderless group exercise, and business games. ACs generally operate under a trait paradigm, whereby the aim of the AC is to obtain measures of stable dimensions that an individual will display across the different exercises. Usually, a dimension must be assessed at least twice during the AC (Cook, 1998), and will commonly be assessed between 3 and 4 times (Ballantyne & Povah, 1995). Put another way, the objective of the AC is to obtain a measure of an individual's dispositional characteristics or the stable characteristics inherent within an individual (Fleenor, 1996) as opposed to situationally-specific behavioural responses to individual exercises.

Cook (1998), Ballantyne and Povah (1995), Thornton (1992), and Thornton and Byham (1982) describe the content of ACs as including exercises that may have come 'off the shelf' (i.e., generic exercises), or exercises that have been designed specifically for a particular job. Such exercises can be categorised as Group Exercises, Individual Exercises and Written Exercises. Although the summary below is by no means exhaustive, it provides a guide to the different types of exercise that are commonly encountered in modern ACs. The selection of a particular form of

exercise should be based on its relevance to the requirements of the target job. Due to considerable variation across job requirements, it is likely that new forms of AC exercises will need to be constructed accordingly. Such exercises will need to be guided by, and congruent with, job analysis outcomes.

Group Exercises

Revealed difference technique discussions: A discussion of an individual's priorities in a contrived scenario, relative to the priorities of the group as a whole.

Leaderless group discussions: A group discusses a point of contention or general topic, and no leader is appointed to regulate the group.

Assigned role exercises: Individuals are allocated particular roles and are provided with a scenario where they must compete for their share of a budget, etc.

Group Discussion Exercises: This can be either co-operative, or competitive. In the former, group members discuss an issue or solve a problem as a team, and in the latter, group members are required to persuade or negotiate on an issue.

Command Exercises: Involves individuals solving a practical problem, e.g., constructing or completing a structure.

Business Simulations: Often involves the use of computers and rapid decisions with incomplete information, under inconsistent conditions.

Team Exercises: Two teams advocate for opposing viewpoints.

Individual Exercises

Presentation Exercises or Lecturettes: Individuals prepare for an oral presentation. The level of formality of the presentation is predetermined by the requirements of the

target job. Presentation exercises are often used in ACs for positions in sales, marketing or training.

Role Plays: The individual is required to adopt a role related to those required in the target job. For example, a disgruntled customer or a person with a grievance.

Interview Simulation Exercises: These involve one-to-one interactions with a candidate and an interviewer who plays the role of a customer, subordinate or other role, depending on the requirements of the target job.

Fact-Finding and Decision-Making Exercises: These exercises purport to assess the analytical skill of candidates. Such an exercise could include a situation where a candidate has been requested to formulate sound and cogent arguments surrounding an issue of contention.

Written Exercises

In-Basket Exercises: A simulation exercise, where candidates are given a set of papers to action (i.e., memos, reports, junk mail, letters, etc) that mimic those that may be encountered in the target job. In-Baskets are supposedly useful for assessing an individual's capacity to manage several issues at one time, delegation, planning, organising and being able to prioritise items of more or less importance.

Analysis Exercises: In these exercises, candidates are required to review information that is presented either numerically or verbally, and to submit a report on the basis of this information. Written and oral communication are of particular salience in analysis exercises.

Biographies: Candidates are required to write a biography for such events as, for example, an obituary for the candidate's own death. Once again, written or oral communication skills are of interest. The extent or type of information assessed in

this, and any, type of exercise (e.g., facial expression, handwriting neatness, etc) should ideally depend on the outcomes of job analysis data.

The Trait Paradigm

The trait paradigm is the theoretical foundation under which most ACs operate (Klimoski & Brickner, 1987). Trait theories emphasise the notion that individuals possess particular and consistent attributes and dispositional characteristics that distinguish one person from another (Byham, 1970). At the core of this definition of the trait paradigm of human attributes lays the notion that behaviour should be reasonably consistent across situations. Stability over time and space is noted as being a "prerequisite of trait validity" (Matthews & Deary, 1998, p. 50). Under the trait paradigm in the AC context, if an individual is rated highly on any given specific dimension on one exercise (e.g. problem solving), then that individual should score (relatively) highly on that dimension on subsequent exercises (Fleenor, 1996).

Howard (1997) described the characteristics measured in ACs as being composed of mixed collections of "traits (e.g., energy), learned skills (planning), readily demonstrable behaviors (oral communication), basic abilities (mental ability), attitudes (social objectivity), motives (need for achievement), or knowledge (industry knowledge) and other attributes or behaviors" (p. 22). It appears that all of the characteristics mentioned, except perhaps for 'behaviours', describe notions that are reminiscent of the trait paradigm, in so much as they represent attributes that are theoretically assumed to hold some cross-situational stability (at least over the course of an AC). In practice, the term trait often tends to be associated with personality variables. Frequent terms that are used to describe these characteristics, whilst avoiding connotations associated with personality theory, include 'dimension' (e.g.,

Lievens & Klimoski, 2001) and 'competency' (Spencer and Spencer, 1993). Also, particularly in measurement theory and psychometrics, an underlying trait is referred to as a latent construct. By and large these terms are considered as interchangeable in the present study, as descriptions of characteristics that are assumed to hold relatively enduring cross-situational consistency. However, given the subtle differences between the terminologies outlined above, certain terms may be more appropriate than others given particular contexts, and will be referred to accordingly in this dissertation. Sackett and Dreher (1982) identified the first empirical evidence against the use of the trait paradigm in the AC context, where their results from an AC study did not show evidence for the dimensions intended to have been measured.

Construct Validity and the Exercise Effect

Construct validity concerns the extent to which a hypothetical trait or construct is measured via a form of psychological test (Anastasi & Urbina, 1997). Tenopyr (1977) argued that the various forms of validation (e.g., criterion validation, content validation, known groups validation) are fundamentally all evidence for construct validation. For the purposes of the present study, as for numerous studies on ACs, it is useful to divide the holistic concept 'construct validation' into its various evidential components. As evidence of construct validity for a given trait or construct, Campbell and Fiske (1959) proposed the concepts of convergent and discriminant validity. The former of these notions implies that constructs that should theoretically be related to one another should accordingly correlate with one another. The latter notion suggests that theoretically distinct constructs should remain unrelated to one another. The authors suggested that in order to research these properties of a given psychological trait, a multitrait-multimethod matrix should be developed, whereby multiple

measurements of traits are obtained, under the utilisation of multiple methods by which to investigate these traits. To evaluate the research on assessment centres using these principles, it is also important to realise that Campbell and Fiske (1959) considered both convergent and discriminant validity to be complementary, and so it would seem, comparative forms of construct evidence. As the authors state, "for the establishment of construct validity, discriminant as well as convergent validation is required" (Campbell & Fiske, 1959, p. 81). Particular attention should be drawn to the words 'as well as' in this quotation, which imply that both sources of evidence should be considered together for an acceptable method of construct validation.

Campbell and Fiske (1959) suggested that for acceptable construct validation, measurements of the same trait across different methods (monotrait-heteromethod measurements, i.e., validity coefficients) should be higher than both different traits measured across different methods (heterotrait-heteromethod measurements) and, particularly relevant in the AC context, different traits measured via the same method (heterotrait-monomethod). The AC process fundamentally functions by way of a multitrait-multimethod mechanism. By definition, ACs employ multiple methods by which to measure multiple traits (Carrick & Williams, 1999) and as such, the AC presents itself as a multitrait-multimethod psychological measure. To achieve acceptable construct validity, according to Campbell and Fiske, ACs need to display higher monotrait-heteromethod correlations than either heterotrait-heteromethod or heterotrait-monomethod correlations.

Through factor analysis, Sackett and Dreher (1982) investigated the construct validity of ACs in an effort to examine the extent to which three ACs actually measured the dimensions that they were designed to measure. In all three ACs, the within exercise ratings correlated more highly than dimensions measured across

exercises, with factor patterns clearly representing exercises (heterotrait-monomethod correlations) rather than the intended dimensions (monotrait-heteromethod correlations). This study provided no support for the view that ACs measure complex constructs such as leadership or problem solving. Although criterion validity evidence for ACs may exist, the results found in Sackett and Dreher and other studies since (see Hough & Oswald, 2000) raise questions as to why and how ACs are predictive of managerial success. As Sackett and Dreher pointed out, the notion that ACs do not appear to measure the constructs that they were intended to measure does not render the criterion-related evidence in favour of the process invalid. It does, however, suggest that the justification for inferences made about ratings in the AC on the basis of traditional content validated methodology may, in itself, lead to erroneous conclusions.

Subsequent studies on AC construct validity have consistently shown the pattern of 'same dimensions across different exercises' exhibiting low correlations (a lack of evidence for convergent validity), and ratings of 'different dimensions within the same exercise' being highly correlated, (a lack of evidence for discriminant validity). Taken together, these findings have given rise to what has been labelled 'the exercise effect†' (Bycio, Alvares & Hahn, 1987; Carrick & Williams, 1999; Chan, 1996; Fleenor, 1996; Jones, Herriot, Long & Drakeley, 1991; Joyce, Thayer & Pond, 1994; Russell, 1987; Lievens, 2002; Robertson, Gratton & Sharpley, 1987; Silverman, Dalessio, Woods & Johnson, 1986; Spector, 2000; Turnage & Muchinsky, 1982; Turnage & Muchinsky, 1984). These studies and summaries suggest that ACs are not working as they were designed or intended to, in that they do

† Note: The 'exercise effect', in this context, is a method effect and should not be confused with the main effect for exercises in ANOVA terminology. The AC exercise effect reflects the variation of the performance of individuals within exercises as opposed to behaviour measured across traits or dimensions. In ANOVA, the 'AC exercise effect' is reflected in the interaction term between the object of measurement and the main effect for exercises. Specifically, the effect for persons crossed with the effect for exercises. This supposed method effect is only regarded as a source of measurement error under the traditional dimension-specific AC model, which attempts to measure trait-based variables.

not appear to measure the dimensional characteristics that they purport to. Table 1 shows the average convergent and discriminant validity coefficients for a sample of studies showing the exercise effect. Most of the studies mentioned show higher correlations between different traits measured within an exercise than measurements of the same trait sampled across exercises. None of these studies show a particularly convincing picture that variables are being assessed in ACs that hold stable characteristics across exercises. Various attempts have been made to increase the evidence for construct validity of ACs through various methods, as described in the next section.

The Importance of Construct Validation in ACs

It could be argued that, given the predictive power of the AC process, the discrepancies in construct validity represent an issue of little importance. The contention is often presented that ACs gain their predictive properties from content validity, in that they theoretically sample a wide range of the factors that comprise target jobs (Neidig & Neidig, 1984; Norton, 1977; 1981; Sackett & Dreher, 1982; 1984; Thornton, 1992). Lowry (1996) however, suggested that content validation is insufficient when purporting that trait-based variables are measured in ACs. In this view, the aim of the AC is not merely to sample the domain of what comprises a particular job, but goes over and above this to infer the measurement of psychological traits. If the justification for trait measurement is derived, properly, from a job analysis, then ignoring the validity of subsequent trait measures is akin to neglecting the trait requirements of the job.

Table 1

Average Convergent and Discriminant Validity Coefficients of Assessor Ratings in a Sample of AC Studies

Source	Different traits within exercises	Same traits across exercises
Sackett & Dreher (1982)		
Company A (n = 86)	.64	.07
Company B (n = 311)	.40	.11
Company C (n = 162)	.65	.51
Turnage & Muchinsky (1982)		
Sample A (n = 1,028)	.53	.45
Sample B (n = 1,028)	.52	.44
Silverman, et al. (1986)		
Sample A (n = 169)	.65	.54
Sample B (n = 178)	.68	.37
Russell (1987) (N = 75)	.53	.25
Bycio et al. (1987) (N = 1,170)	.75	.36
Robertson et al. (1987)		
Organisation 1 (n = 41)	.64	.28
Organisation 2 (n = 48)	.66	.26
Organisation 3 (n = 84)	.60	.23
Organisation 4 (n = 49)	.49	.11
Lievens & Conway, (2001)		
Average across 34 studies	.34	.34

Source: Adapted from Reilly et al., (1990).

In terms of convergent and discriminant validity, the robust finding with many of the studies performed on ACs is that they measure exercise constructs (i.e., performance on individual exercises), yet over and above this, the evidence suggests that any inference of traits may be erroneous (Lowry, 1995). To elaborate, the questions surround what the construct validity evidence, taken holistically, is suggesting about the constructs that are being measured in ACs. This, in itself, is an inherently important reason why construct validation is a variable of consequence to the discipline of industrial/organisational psychology. Making claims with no empirical evidence could threaten the credibility of a practice or even a discipline in its own right.

Further to this, Lowry (1996) and Norton (1977) have suggested that construct validity may be a moot point for litigation against companies attempting to use AC technology, particularly for the purposes of recruitment and selection. Both authors contend that courts in the United States recognise construct validation procedures. As construct validation is arguably the most fundamental type of validation in assessment procedures (Tenopyr, 1977), it is highly conceivable that without empirical construct validation through evidence of the discriminant and convergent validity of an AC, organisations may expose themselves to court cases that may be costly in terms of monetary losses, as well as being threatening to their credibility and reputation. Thus, under this argument, it is with some urgency that the lack of construct validity evidenced in ACs should be investigated and rectified accordingly. The practical significance of this notion extends to the multitude of organisations that use ACs, given the finding that most of these do not assess the validity of their assessment ratings (Spychalski, Quinones, Gaugler & Pohley, 1997).

Sackett and Dreher (1982) examined the construct validity of three ACs, through an analysis of the interrelationships among the dimensional ratings of candidates between and within AC exercises. Sackett and Dreher made the suggestion that construct validation, over and above the inherently job related nature of ACs (i.e., content validity), was essential. This is because firstly, AC dimensions are often complex (e.g., *interpersonal skills, leadership*) which makes an inference to underlying and stable attributes potentially more difficult. Secondly, the AC process is, in itself, composed of several complex inferential steps that lead to the outcome of making a prediction of managerial potential. These steps involve the observation of candidates in AC exercises, recording numerous trait ratings for candidates, to the final step of producing overall predictions of managerial success. Thus, the underlying traits that are involved in becoming a successful manager perform a vital role in the AC process. Kudisch, Ladd and Dobbins (1997) argued that providing traditional dimensional feedback to the participants of ACs, with the implication that they are in fact measuring such dimensions, is highly problematic. In the developmental context, the low convergent and discriminant validity issue has dire implications in terms of leading to erroneous feedback. That is, participants may be provided feedback on the basis of stable ability traits that are not actually being measured (Fleenor, 1996). Thus, decisions in organisations can be falsely made on the basis of AC measurement.

Construct Validation of AC Dimensions Through the Nomological Network

Theoretically, evidence of how well a measure defines a construct could be demonstrated though the extent to which that measure fitted legitimately into an arrangement of expected relationships (Nunnally & Bernstein, 1994). Such a pattern of evidence was labelled the *nomological network* by Cronbach and Meehl (1955),

where logic states that to be construct valid, a measure must fit in with the theory relating to a particular construct. Using this logic, a small number of studies have attempted to link overall dimension ratings in ACs to ratings of external measures (such as personality or ability measures) to establish the extent to which such ratings formed an expected pattern of construct relationships.

Mixed evidence has been reported for the extent to which AC dimensions fit within a nomological network. Shore, Thornton, and Shore (1990) and Thornton, Tziner, Dahan, Clevenger, and Meir (1997) classified psychological (cognitive ability and personality) measures as being either related or unrelated to specific AC dimensions. Generally small expected relationships were found with cognitive-ability-related and performance-related-dimensions. Further convergent and discriminant validity evidence, in the form of small correlations, was also found for interpersonal effectiveness dimensions. Chan (1996) and Fleenor (1996) found evidence against the notion of AC dimensions showing construct related evidence through the nomological network. They reported a lack of convergent and discriminant relationships with nearly all of the expected associations.

Factors that May Improve AC Construct Validity: The Limited Information-Processing Model.

Lievens and Klimoski (2001) posited two models that sought to explain why ACs lack construct validity. The first was the *limited information-processing model*, which suggests that since assessors have restricted information-processing capabilities, they are not necessarily able to fulfil the requirements demanded for trait-based measurement (e.g., Miller, 1956). The techniques for minimising this potential problem are discussed in the following, and incorporate suggestions from Lievens

(1998) and other authors who have investigated these suggestions empirically. Lievens (1998) reviewed the factors which research suggests are implicated in the facilitation of the construct validity of AC ratings. Twenty-one studies conducted between 1976 and 1997 were included in the review. Based on his review, Lievens suggested several techniques for maximising the validity of AC ratings. Many of these suggestions parallel those proposed by Ahmed, Payne, and Whiddett (1997), who interviewed experts and practitioners with regard to best practice in the design of AC exercises. It was suggested that a small number of dimensions be used when assessing candidates. Also, dimensions should be used that are conceptually distinct from one another so as to maximise ease of discrimination for assessors, and dimensions should be defined in a concrete and job related way. In terms of the situational exercises used, exercises should be selected on the basis of their ability to elicit behaviours that are relevant to the dimensions being measured. Any role players used should be trained and their actions should be standardised to the greatest degree possible, so as to avoid across-exercise variance. Moreover, it was advised that role players should seek to elicit dimension-related behaviours, and that dimensions should be revealed to the assessees, especially in developmental ACs.

The suggestion has been made that assessors should be provided with observational aids (e.g., behavioural checklists and video recordings of the AC), and that dimensions should be included in a checklist, utilising a maximum of 12 behaviours for each dimension. Checklist behaviours should be grouped in naturally occurring clusters. Rating biases should be minimised by using a rotation system, whereby assessor/assessee combinations are rotated over the course of the AC. This would represent a rudimentary form of randomisation in which rater idiosyncrasies would be systematically allocated among the pool of assessees. The suggestion has

been made that one dimension only should be rated across exercises (Robie, Adams, Osburn, Morris & Etchegaray, 2000). Some empirical work has been performed with respect to these suggestions to investigate their relative efficacy. In general, these studies conclude that careful consideration given to particular design characteristics associated with ACs is important in terms of maximising evidence for construct validity (Lievens & Van Keer, 2001). It should be noted, however, that some studies that have purposely implemented some of these general suggestions have not been successful in terms of increasing evidence for the measurement of trait-based constructs (Chan, 1996; Schneider & Schmitt, 1992; Fleenor, 1996). The findings from specific studies that have investigated the limited information-processing model are summarised in the following.

Rating Dimensions Subsequent to Agreeing Upon Dimensional Ratings

The exercise effect could possibly be influenced by the fact that assessment dimensions are often rated directly after or during an exercise. This might have an impact on the extent to which AC exercises are treated individually, or whether assessors look for behavioural patterns across different exercises. Silverman, et al. (1986) made great efforts to try to minimise the exercise effect. The authors argued that, traditionally, assessors in an AC are required to rate each dimension on each exercise. Given this, assessors may be forced to process rating information in terms of exercises, thus the exercise effect prevails as assessors rate in terms of exercises. Silverman et al. arranged their AC in such a way that the assessors were asked to rate candidates on dimensions after the rating for each dimension had been agreed upon by a group of assessors. The authors labelled this approach the 'within-dimension' method, in which the assessor group firstly observed and recorded behavioural

information across all exercises. Secondly, the information recorded for each of the dimensions in the AC was displayed on an overhead and was discussed. Thirdly, from the information gleaned in this discussion, the assessors rated overall dimension ratings for each candidate. Fourthly, these ratings were discussed so as to reach an agreement for each dimension assessed for each candidate. Fifthly, assessors privately rated the performance of each dimension for each candidate on each exercise. Such a procedure appears prolonged, and may even detract from the original behaviour assessed.

The emphasis in the Silverman et al. (1986) study was supposedly on performance related to dimensional characteristics, as opposed to exercise performance. This approach was similar to that used in the original AT&T AC where dimensions were only rated after all of the exercises had been observed, that is, dimensions were not rated after or during each exercise. Despite modifications being made, Silverman et al. found that the pervasive exercise effect prevailed, with factor analyses not revealing any clear dimensional clusters. Silverman et al. interpreted factor rotations of assessor ratings as revealing clear exercise effects. As Silverman et al. commented, "the within-dimension method still showed a smaller amount of discriminant validly than anticipated, and it also showed source or exercise variation" (p. 575).

Harris, Becker, and Smith (1993) argued that the Silverman et al. (1986) study was rendered problematic by the notion that requiring assessors to determine an overall rating first may have artificially forced assessors to be more consistent in their ratings of dimensions, even though, in reality, the Silverman et al. study did not actually show strong evidence for the measurement of trait-based variables. In a systematic replication of the Silverman et al. study, Harris et al. sought to investigate

the utility or the improvement in construct validity due to deriving overall dimension ratings only after the dimension rating for each exercise was determined. They found no increase in cross-situational consistency in terms of dimensional characteristics. The more common finding of different dimensions within the same exercise correlating more highly than the same dimension across different exercises was observed.

Having Assessors Rate A Singular Dimension Across Exercises

A curious approach towards maximising the possibility that assessors will display ratings that represent trait-based variables was suggested by Adams (1997). Adams hypothesised that exercise effects occurred because the same raters were measuring different dimensions within a given exercise. By having each individual assessor rate a single, particular dimension across the different exercises, Adams found evidence for the measurement of dimensions across exercises. This same premise was repeated using more robust methodology by Robie, et al. (2000) with almost identical results.

The implications of the principles presented above, present some practical considerations in that as many assessors need to be employed as there are assessment dimensions being assessed for each individual. More serious though, is the notion that these studies are in danger of conveying that the extent of correlation between the variables measuring a dimension is more contingent on the rater, and less so on the presentation of a behaviourally manifest trait-based variable. If it were trait-based variables that were being shown, and the raters held a shared understanding of what defined a dimension behaviourally, then it should not matter which assessor rated which dimension. Trait-based characteristics should, theoretically, remain relatively

stable and enduring across exercises. The underlying premise in the above studies really creates a kind of 'mono-trait effect' across a singular rater. In this sense, the measurement of a singular trait is confined to a singular 'method'. In this case that method is a rater. This does not necessarily mean that the method effect has gone, but rather, it may have been turned on to dimensions instead of exercises.

Reducing Cognitive Load On Assessors and Organising Ratings

Some evidence suggests that a lack of construct validity in AC ratings might be influenced by the sheer demands of the arduous task that assessors face. The suggestion related to this involves reducing the demands on assessors, so as to potentially increase the quality of the assessment. Jones et al. (1991) attempted to improve the validity of an AC for the selection of naval officers by introducing procedural changes. The number of dimensions being assessed was reduced to minimise the cognitive demands put upon assessors. Within a document designed to assist the assessors to organise rating evidence, predictor information was organised under each relevant dimension on a schedule, which visually indicated the relative predictive validity of each dimension. Discussions at the end of the assessment process were focused on areas of disagreement or disparity. Following the modifications, the study found no improvements in predictive validity for training criteria, although there were improvements in the prediction of voluntary turnover. The study found that despite the alterations, the measurement of individuals' traits or abilities was not obtained through AC ratings, and the exercise effect prevailed. Jones et al. (1991) and Gaugler and Thornton (1989) do, however, conclude that the utilisation of fewer assessment dimensions leads to more accurate ratings, which may

work by reducing the cognitive load that assessors need to contend with when assessing candidates.

Another study by Reilly, Henry, and Smither (1990) investigated the extent to which excess cognitive load may lead to a lack of construct validity. The cognitive load controversy suggests that assessors may be unable to assess the ability traits or dimensional characteristics of assessees, due to the fact that they are overloaded with information. This makes processing a large amount of information unmanageable or impossible (Spector, 2000). Reilly et al. (1990) attempted to investigate the extent to which this phenomenon exacerbated the lack of construct validity associated with the AC process by providing a checklist of 273 behaviours to use in order to rate candidate performance in assessment exercises. This manipulation lead to an improvement in the convergent and discriminant validity of assessor ratings. The use of behaviour checklists increased the average convergent validity coefficient from .24 to .43, and decreased the average discriminant validity coefficient from .47 to .41. The latter coefficients here demonstrated slight improvements in the extent to which same traits were judged as being related and different traits were judged as divergent. Note that one hopes to find relatively high convergent coefficients and relatively low discriminant coefficients for construct evidence. Spector (2000) asserted that these results appeared promising, and suggested that by imposing more structure, and lessening the cognitive load on assessors, construct validity might improve.

In the study mentioned above, Reilly et al. (1990) reported improvements in convergent validity with their use of behavioural checklists to reduce the cognitive demands put upon assessors. The study only found a difference in overall correlation of .06 in the reduction of discriminant validity. There are empirical questions surrounding whether this is, in actual fact, a notable improvement in the construct

validity of ACs. Viewed from an alternative perspective, Reilly et al. found only a negligible difference in the extent to which ratings variance was being explained by unrelated constructs, versus related constructs. Note that traditionally, discriminant coefficients should be small, indicating little relationship between theoretically unrelated constructs. Convergent validity coefficients should be comparatively high, indicating strong relationships with theoretically related constructs. In Reilly et al.'s study, convergent validity of the dimensions measured explained 18% of the variance in assessor ratings. Evidence for discriminant validity is manifest in smaller discriminant than convergent correlations. However, in this study, discriminant correlations explained 17% of the variance in assessor ratings. Viewed in this way, this results in a somewhat redundant set of information in regard to what was actually being measured in the AC. The author's results suggest there is almost as much construct validity in their AC as there is no construct validity, as evidenced by the similar convergent and discriminant correlations. Reilly et al. endeavoured to reduce cognitive load on assessors through the use of behavioural checklists. However, their results showed little difference in terms of explanatory variance in ratings being accounted for by exercise effects or the measurement of dimensions. The contention as to whether or not this AC was actually measuring ability traits is, by the very nature of the convergent and discriminant validity coefficients observed, seemingly unresolvable.

Subsequent researchers have also proposed the use of behavioural checklists to decrease cognitive load. This notion was investigated by Brannick, Michaels, and Baker (1989). These authors utilised behavioural checklists to organise ratings across two in-basket exercises. Despite these modifications to the AC process, very little cross-situational consistency in dimensional ratings was reported. This was surprising

because Harris et al. (1993) suggested that this study should have provided optimal conditions for cross-situational consistency due to the fact that Brannick et al. used different versions of the same exercise, behavioural checklists, and there was no time limitation for assessor decisions and for the completion of documentation pertaining to candidate behaviours. Donahue, Truxillo, Cornwell & Gerrity (1997) also found very similar effects for the use of behavioural checklists, with negligible differences shown between average convergent and discriminant validity coefficients. In a more recent summary of the literature on the use of behavioural checklists, Lievens and Conway (2001) found no average difference between the proportions of dimension variance found for 34 studies that either used, or did not use, behavioural checklists.

The number of dimensions that should be assessed in an AC has also lead to some debate about lessening the cognitive load upon assessors. Lievens and Klimoski (2001) argued that no more than four or five dimensions should be included within one particular exercise. Empirical evidence was found to support this suggestion recently. In a study of 34 different ACs, Lievens and Conway (2001) found that when five or fewer dimensions were used, variance in scores could be better explained by dimension variance. The results of this study were similar to those found previously by Gaugler and Thornton (1989), who suggested that the optimal number of dimensions that should be assessed across an AC lies between five and seven (also, see Chan, 1996).

Cognitive load considerations need to be weighed against the specifications set for the robustness of a measurement tool. Singular item measures are an issue of controversy, except perhaps for highly deliberated constructs, such as job satisfaction, and present obvious difficulties when considering construct coverage and the calculation of inter-item correlations. Two items may yield a reasonably rudimentary

estimate of construct variance through the calculation of only one inter-item correlation. With three measures, the entire construct domain may not be fully covered, however, the third measure allows an additional inter-item correlation for comparison. Theoretically, this is far more robust than one or two measures of a construct. It is therefore argued that in an AC, construct measurements should allow for at least three judgements of a construct across at least three different situations. It may not be that the entire construct domain is necessarily covered by this measurement, and that is why it is vital to select simulations that reflect critical and ecologically valid forums in which a construct might be manifest.

The reduction of cognitive demands on the basis of the process by which dimensions are assessed have yielded mixed findings reported by Reilly et al. (1990), Jones et al. (1991) and Schneider and Schmitt (1992) discussed later. The approaches used in these studies could be construed as appearing to force a process, which already appears to be efficacious in terms of its relationship with criterion measures, into a paradigm that it does not appear to fit consistently. ACs are often regarded as efficacious predictors of certain criterion measures, so it appears logical to investigate why they work, and under which paradigm they work. Perhaps forcing them into a trait paradigm actually undermines their predictive validity. It might also inhibit attempts to establish an understanding of the true nature of the process by which ACs predict promotion, performance and related measures. In a similar vein, a lack of understanding as to why ACs predict criterion measures may deter the ability to improve the AC process to any significant degree.

The Use Of Video Recordings

Video recordings have also been investigated in terms of their propensity to facilitate trait-based measurement. Buckner (1984) found preliminary evidence that there was very little difference between ratings derived from live assessment and from a video recording. Ryan, Daum, Bauman, Grisez, Mattimore, Nalodka, and McCormick (1995) also compared the ratings obtained from direct observation against those obtained from video-based observation in an AC context. The authors found negligible differences in terms of rating and observation accuracy when comparing live and videoed assessments. The authors conveyed; "These results suggest that videotaping is not worthwhile if the purpose is to enhance accuracy in rating and observation" (p. 668). Allowing assessors to rewind and re-observe behaviour through a video did, however, lead to greater specificity in the recording of behavioural notes during the AC.

Dimensional Transparency

Some research has aimed to investigate how an awareness of the dimensions being assessed, on the part of the participants, might maximise the potential for evidence of construct valid ratings. Kleinmann (1993) argued that the dimensions being assessed in ACs are not transparent enough for assessees, and that this lack of recognition of assessment dimensions may account for the lack of convergent and discriminant validity inherent in the AC process. It was argued that assessees may be unaware of the rating dimensions, and may be unaware of the behaviours that are relevant for each dimension being assessed. In Kleinmann's study, candidates recorded their perceptions as to which dimensions they thought they were being assessed under. It was found that candidates who more accurately recognised the

assessment dimensions tended to perform better on the AC. Convergent validity was found to improve when candidates accurately perceived that the same dimension would be assessed across two different exercises. Implications for the usefulness of these findings in actual AC practice are controversial.

Questions surround what characterises a person who is better able to recognise the dimensions under which they are being assessed. Perhaps these individuals are better able to recognise policy-capturing factors, or factors that may please a given audience, and thus, such an ability assists them to get ahead in an organisation. This might lead to the measurement of a latent construct akin to impression management skill, which is discussed later. Kleinmann also suggested that assessees should be told about the assessment dimensions that they are being rated on prior to the assessment process. This could, however, change altogether what it is that ACs measure.

One possibility is that ACs may measure latent constructs (Chan, 1996) that have not yet been identified, and some evidence suggests that these constructs could be useful in predicting managerial potential (Ballantyne & Povah, 1995). If assessees are informed of the assessment dimensions, the AC process may turn into a test of memory ability or impression management skill. Those candidates who are better able to remember the dimensions that are being assessed or who are better able to manage their impressions may be more likely to perform better on an AC. There may be unidentified and potentially dire consequences associated with interfering with the subtle workings of the AC process in this way. It seems that such a method would be aiming to force the AC process into a paradigm that has questionable relevance in this context. In any case, it seems unlikely that making the dimensions transparent will lead to any substantial gains in the measurement of trait-based constructs. Lievens and Conway (2001) effectively found no difference between the proportions of

variance in AC ratings explained by dimensions among 34 studies that either used or did not use transparent dimensions.

Exercise Transparency and Opportunities To Express Behaviour

The extent to which candidates are familiar with particular exercises has been suggested as a factor that may influence the quality of measurement in AC ratings. The proposition has also been made that candidates should be given the opportunity to manifest the relevant behaviours being assessed in the AC. Woodruffe (1993) made several suggestions with regard to factors that may act to decrease the occurrence of the exercise effect in assessor ratings. AC architects should begin, in this view, by making sure that dimensions can be assessed independently in each exercise. Dimensions should also be articulated to assessors in such a way that they are easily understandable, well defined, and easily discriminable from each other. Woodruffe also suggested that the extent of the exercise effect was a function of the degree to which candidates were familiar with different assessment exercises. However, as with the previous argument pertaining to familiarising candidates with assessment dimensions, such familiarisation may actually change what it is that ACs measure, in terms of any valuable latent, or other, constructs that may be assessed.

Woodruffe (1993) suggested that candidates might not always be given the opportunity to express the behaviours necessary for assessors to make inferences regarding the targeted constructs. Implications from this suggestion might arise in situations where ACs attempt to measure too many dimensions relative to the number of exercises. Turnage and Muchinsky's (1982) study, for example, involved an assessment of all dimensions rated in each exercise employed in their AC. This is acceptably, although not frequently, observed in practice. In such cases where all

dimensions are rated in each exercise, Woodruffe (1993) argued that candidates might not have sufficient opportunities to reveal all the behaviours that relate to appropriate dimensions. However, this argument does not provide an adequate explanation of the clear factor loadings that revealed exercise effects in Turnage and Muchinsky's (1982) study. Even if candidates were not able to reveal all of the behaviours that were related to dimensions in all of the exercises, if this were the only reason for the exercise effect, then it might be expected that at least some of the dimensions across exercises may be related to one another (Robertson, et al., 1987).

Form and Content of AC Exercises

Theoretically, different types of AC exercise, and the content of those exercises, have the potential to influence the quality of the ratings produced in ACs. Several authors have suggested these might be important considerations for AC design, particularly Schneider and Schmitt (1992), who suggested that the form and content of an AC might be important determinants of the extent to which they yield construct valid ratings. In an effort to reduce the exercise effect, Schneider and Schmitt ensured that assessors were highly familiar with dimensions through thorough training, and also reduced the cognitive load on assessors by minimising the number of dimensions being assessed. An exercise effect was still found despite these precautions. Additionally, candidates appeared to vary in performance across different forms of exercise. In this study, there were two major forms of exercise employed: group discussions and role-play exercises. Schneider and Schmitt suggested that the form of an AC exercise is of great importance in explaining the variance in ratings, and as such the form of AC exercises should reflect those situations that will be encountered on the job.

Lievens (2002) also suggested that design considerations in ACs are integral influences with respect to construct validation. In an experimental simulation of an AC, Lievens presented assessors with videotaped candidate performances that varied according to the consistency of candidate behaviour across exercises and across dimensions. Evidence for convergent validity was found when videotaped candidates were presented who appeared to perform consistently across exercises. Evidence for discriminant validity was found when videotaped candidates performed differentially on different dimensions. Lievens reported that the efficacy of the dimensions in this simulated AC was, at least in part, influenced by design characteristics. Specifically, the choice of assessors was seen as important as, in this study, the use of student assessors reduced the extent to which ratings reflected trait-based constructs. Lievens (2002) also suggested, from the results of this study, that the absence behavioural checklists, the use of too many dimensions and the use of very different exercises may undermine construct validation. However, certain organisations might well be justified in selecting a range of different exercises for assessment in order to sample the focal job as holistically as possible. Also, with fewer dimensions, perhaps a wealth of information is lost at the price of maximising convergent and discriminant validity. Moreover, and according to Lievens, more importantly (p. 683), the results of the study suggested that AC evaluations are influenced by trans-situational candidate behaviour. Thus, it was found that the behaviour of candidates across situations had a profound effect on the way in which candidates were rated.

Factors that May Improve AC Construct Validity: The Expert Assessor Model.

Lievens and Klimoski (2001) proposed a second model to explain why ACs lack construct validity. Namely, the *expert assessor model.* According to this model,

the quality of assessment ratings is contingent on whether an assessor is a novice or an expert. The underlying notion here is that an expert assessor will hold firmly established mental models (schemata), which they can apply in their assessment of individuals. As Lievens (2002) reflects, such cognitive frameworks are theoretically adaptive in terms of the extent to which they direct attention, recall, categorisation, organisation, and integration processes. Experts can develop their schema by abstracting from their education, training, and experience. In this respect, raters with a postgraduate degree in psychology would be expected to yield construct valid ratings. Novices, under this model, are not expected to hold the mental structures necessary to yield construct valid ratings. Of particular interest in the AC literature are the small number of studies that have looked at the effects of frame of reference training, and the effects of employing psychologists as assessors.

Frame Of Reference Training

Frame of reference (FOR) training has received much attention in the recent literature as a potential vehicle with which to maximise the measurement of traits in ACs. Lievens (1998) made suggestions with regard to the make up of the assessor panel in ACs. He proposed a focus on the quality of training provided to assessors rather than the length, and further that the principles of FOR training should be included in the training scheme. FOR procedures represent a form of standard setting for raters that encourages and facilitates the construction of shared mental models with regard to varying levels of performance (Lievens, 2001a; Schleicher & Day, 1998; Sulsky & Day, 1992). Two studies in particular reported that better quality, in terms of measuring trait-based variables, in ACs was obtained through the application

of FOR training procedures (Lievens, 2001a; Schleicher, Day, Mayes & Riggio, 2002).

In the Schleicher et al. (2002) study, interrater reliability was found to be consistently higher for a group of student assessors who received FOR training when compared to control assessors. When looking at evidence for the measurement of trait-based variables, however, the results were less clear. Although predictions about convergent validity were not included in the hypotheses of the study, the averaged transformed monotrait-heteromethod correlations were significantly higher ($p < .05$) for the control group (.48) than for the FOR group (.34). This presents evidence that the control group made greater connections between the same traits measured across different exercises than did the FOR trained group, which is counter to expectation. Transformed heterotrait-monomethod correlations (i.e., discriminant validity evidence) were, however, lower for the FOR group (.66) than for controls (.74), which was expected, as discriminant validity coefficients should be relatively small for trait evidence. However, a difference in the discriminant validity coefficient of .08 between FOR trained assessors and controls appears to be marginal in terms of overall magnitude. Overall, the discriminant coefficients appeared to be appreciably higher than the convergent coefficients in this study. This suggests that there were greater correlations in general between different traits than between same traits, regardless of whether an individual was FOR trained or not. This finding is counter to what would normally be expected according to the fundamental notions of construct validation proposed by Campbell and Fiske (1959), who asserted that convergent and discriminant coefficients present complimentary forms of construct evidence.

Evidence of discriminant validity for FOR training was reported by Lievens (2001a). The mean heterotrait-monomethod correlation was calculated as .17 for the

FOR assessors, .24 for the data-driven assessors and .39 for the control group. The data-driven assessor training program in this study encouraged assessors to observe, record, classify and evaluate behaviour. Heterotrait-heteromethod and monotrait-heteromethod correlations were not available, as only one assessment exercise was used in the study and thus important comparative evidence for trait-based measurement was not attainable. Without comparison with convergent validity coefficients, construct inference becomes hazardous. The heterotrait-monomethod findings in this study concurred with generalizability analyses, which found variance components that reflected the interaction effect of candidates and dimensions estimated as being marginally higher for FOR trained assessors than for data-driven and control assessors. In G theory, the interaction between candidates and dimensions acts as evidence for a form of discriminant utility, in so much as it reflects the extent to which the candidates differ in their levels of performance with respect to different dimensions. Again, due to the singular exercise under scrutiny, evidence for convergent validity was unattainable.

In the Schleicher et al. (2002) study, average heterotrait-heteromethod correlations were reported as lower for FOR trained assessors (.28) than for controls (.48), which was expected as FOR trained assessors should be able to discriminate better between different traits assessed across different exercises. Schleicher et al. (2002) concluded that since the correlations for all three types of assessment were higher for controls than for FOR trained individuals, greater precision was exercised by the FOR group. The evidence in favour of FOR training in this study does not appear to be entirely convincing, however, with a lack of clarity in the evidence for actual construct measurement. This concerns the convergent and discriminant

evidence revealed in the comparison between controls and FOR trained assessors mentioned earlier.

Schleicher et al. (2002) and Lievens (2001a) also investigated the extent to which FOR trained and non-FOR trained assessors differed in the accuracy of their ratings as indexed on Cronbach's (1955) accuracy indices, and Borman's (1977) differential accuracy index (BDA) (see also Sulsky & Balzer, 1988). In the Schleicher et al. (2002) study, rater type explained the following proportions of variance in the four Cronbach indices: elevation ($R^2 = .14, p < .05$), differential elevation ($R^2 = .09, p < .05$), differential accuracy ($R^2 = .03, p < .05$), stereotype accuracy ($R^2 = .01$, *ns*) and the BDA ($R^2 = .00$, *ns*). Lievens (2001a) used the Cronbach differential accuracy (DA) index and the BDA to investigate accuracy. Small amounts of variance in these accuracy indices were explained by type of training. Effect sizes were calculated through a discriminant analysis for the DA ($\eta^2 = .11$) and for the BDA ($\eta^2 = .09$). In both the Schleicher et al (2002) and the Lievens (2001a) studies, the authors concluded that FOR trained assessors were more accurate in their ratings than controls. FOR training was originally formulated to increase rater accuracy (Bernardin & Buckley, 1981), and as such an increase in accuracy was expected. The amount of variance explained in the accuracy indices investigated in terms of overall magnitude was, however, small.

Schleicher et al. (2002) compared FOR assessors with controls further in the extent to which their AC ratings correlated with conceptually similar psychological variables measured externally to the AC through questionnaires (the nomological network). Some limited evidence was found for expected relationships in this regard. Only three out of eight pairs of expected correlations, pertaining to FOR and control ratings, stood out as significantly different from one another in the expected direction

in a 1-tailed *t* test. FOR ratings were also found to be more predictive of supervisory ratings of performance (.32) than ratings from controls (.21). Although the magnitude of difference between these correlations does not appear great, Schliecher et al. performed a 1-tailed *t* test, which revealed a significant difference ($p < .05$). The authors attributed the improvement in predictive validity to improved construct validity in terms of the ratings from FOR trained assessors. In sum, although there may be some improvement in ratings obtained after FOR training, the gains in the AC context, in terms of overall magnitude, appear minimal within the research to date. Thus, these gains appear to be unconvincing when considering improvements made in trait measurement in ACs.

Employing Psychologists As Assessors

Because traits are assumed to be psychological entities, it seems logical that ratings attempting to tap such phenomena should originate from an expert assessor in the psychological field. Indeed, it could be hypothesised that since psychologists are accustomed to thinking in terms of trait-based variables, they may be more apt towards generalising these mental structures to AC ratings. It may be that assessors, be they managerial staff, supervisory staff, or psychologists, hold divergent schema in regard to what they are assessing. Several authors have suggested that assessors utilise schemata, or their existing mental representations, in their ratings of candidates (Russell, 1987; Harris, et al., 1993). Gaugler et al. (1987) suggested that psychologists make for superior assessors as their background, education, and experience, equip them better than others in terms of being able to observe, record, and rate behaviour. These experiences may extend to a better conceptualisation of what traits are, in terms of being dimensions that exist across situations. Such

experiences and training are likely to develop schemata as to how behaviour should be assessed, and thus it is logical that someone who is trained in psychology would be in a better position to produce data to fit a psychological model of assessment.

Lievens (1998) suggested that psychologists should be included as a key component of the assessment process. Several studies have found that using psychologists as assessors can actually lead to greater predictive validity in AC ratings (Gaugler et al., 1987). Further to yielding improvements in predictive validity, there is some evidence to suggest that using psychologists as assessors may actually increase the extent to which an AC yields construct valid ratings. Sagie and Magnezy (1997) compared the factor structure of the assessor data for psychologists and managers respectively. It was found that while the predetermined assessment dimensions were found in the psychologist's data, manager's ratings had low construct validity. Further evidence for the effect of more dimensional based variance being found in ratings from psychologists was found by Lievens and Conway (2001). Across 34 studies, Lievens and Conway found that the average proportion of variance attributed to dimensions was .27 for managers and .39 for psychologists, a difference that was found to be significant. Kudisch et al. (1997) found some evidence for construct validity in their study of a developmental AC, with factor analyses revealing a mixture of both exercise and dimensional factors. Eight of the 15 assessors who participated in the study were doctoral level psychology graduates. Thus, just over half of the assessor panel were thoroughly trained in psychology. Additionally, Lievens and Conway (2001) found more variance attributable to dimensions as opposed to exercise variance when psychologists were used as the assessor panel.

A recent study that investigated Lievens' suggestions, and which applied generalizability theory (G theory) to the analysis of AC data, found construct valid

AC ratings (Arthur, Woehr & Maldegen, 2000). The overall goal of this study was to examine the way in which guidelines recommended by research and practice could improve the convergent and discriminant validity of trait-based ratings. Several variables were included simultaneously and non-systematically, in such a way that the relative effects of one variable over another could not be partialed out, as Arthur et al. (2000) acknowledged. Both a generalizability study (G study) and confirmatory factor analysis (CFA) reflected that levels of fit in the data were better for dimensions than for exercise factors. No predictive or criterion-related validity coefficients were available. However, the focus of the study was particularly on construct validity in ratings. Arthur et al. concluded that the lack of construct validity found in previous studies was the result of development and implementation factors.

When interpreting the findings of Arthur et al. (2000), it is important to be aware of which data were analysed. The full AC under consideration consisted of four exercises and nine dimensions. Of these, data from three of the exercises and four of the dimensions were included for analysis. This was done in order to achieve a fully crossed design, so that factorial models from G studies and the CFA could run optimally. Thus, only a proportion of the actual data were analysed. This may be problematic in terms of generalisation to the AC as a whole, as less than half of the total data were examined. Note, however, that Arthur et al. replicated their analyses using confirmatory factor analysis and MTMM's with the full dataset included. The results showed similar patterns to those expressed in the G study.

In the Arthur et al. (2000) G study, one of the variance components that was calculated was likely to be problematic. This variance component related to the object of measurement, persons, which was said to be nested within raters. This suggests that over the eight-year period that the AC was run, individual assessors

rated the performance of their own specific groups of participants across all administrations. It seems possible that there was natural attrition amongst the group of assessors over the eight-year period. In addition, in common with the other studies that found construct validity in their ratings, Arthur et al. employed psychologists as an integral component in their assessment process. In their study, 51% of the assessor staff were post Master's and/or Ph.D. level industrial/organisational psychologists.

Lievens (2001b) found additional evidence that psychologists are superior in terms of the extent to which they are able to ascribe trait-based categorisations to manifest behaviour. Lievens (2001b) also employed the use of a G study to the investigation of ratings obtained from psychology majors, 71% of whom had worked in psychological consulting companies or personnel departments, and a group of managers. Comparatively more variance associated with dimension assessment was explained by psychologist assessors (30.9%) than by managers (19.1%). This suggests that psychologists were better able to make differentiations across the different dimensions employed in the AC. Similar results were found in Lievens (2002) where discriminant and convergent validity was more clearly established for psychologists and, interestingly, managers than for students. Note that Lievens (2001b) analysed only half of the total set of dimensions that were included in the actual AC in order to create a fully crossed design. Although this was unavoidable when using a variance components analysis, it may limit generalisation back to the original AC.

Using psychologists as assessors may appear to be adaptive for maximising both the predictive and construct validity of AC ratings, however there are some issues of contention that require deliberation when choosing psychologist assessors as opposed to employing managerial or supervisory assessors. In terms of predictive

validity, Gaugler et al.'s (1987) meta analysis of AC validity is the study most commonly cited in asserting that manager's ratings tend to have less predictive validity than psychologists' ratings. It is constructive to consider the composition of assessor panels in Gaugler et al.'s various studies. Sixty-four percent of the studies employed managers as assessors, 20% involved a combination of both psychologists and managers, and only 10% of the studies employed psychologists alone. Consequently, it may be that managerial assessors were over-represented and psychologist assessors were under-represented in the Gaugler et al. meta analysis.

Moreover, as conveyed earlier, Fleenor (1996) suggested that as there is a lack of standardisation in the design and execution of ACs, meta-analyses might not be the ideal means by which to judge AC validity. Standardisation, as a potentially contentious issue in AC methodology, was even mentioned in some of the first works on the process (Bray & Grant, 1966). Gaugler et al. (1987) did not include standardisation as a criterion by which studies could be selected into their analysis. As Chow (1996) states "conceptual rigour or research quality is not deemed important in meta-analysis" (p. 110), and adds further, "It is not clear how it is possible to come up with valid information if pieces of invalid information are integrated in the meta-analytic manner" (p. 110). Thus, particularly in light of the issues surrounding standardisation concerning the construction of ACs, researchers using meta-analyses of the AC approach need to ascertain whether or not they are actually comparing like ACs with like ACs.

Perusal of the descriptive statistics in the Gaugler et al. (1987) meta-analysis suggests that there could be practical problems associated with the employment of psychologists as assessors. Of the ACs selected into the Gaugler et al. meta-analysis, most (64%) employed managers alone as assessors. Lowry (1996) performed a

comprehensive survey of AC use in the public sector in the United States. None of the respondents from this sample indicated that they specifically employed psychologists as assessors. The Spychalski et al. (1997) extensive survey of AC practices in the United States further revealed that both public and private sector organisations were unlikely to employ psychologist assessors. Again, similar results were obtained, where the greatest proportion of assessors were line managers (approximately 49%) or staff managers (approximately 26%). Only 6% of cases reported the utilisation of psychologists as assessors. Thus, there may be several well-justified practical reasons that organisations are reluctant to employ psychologist assessors. The costs associated with employing psychologists in this manner, as opposed to managers, are likely to be far greater. Organisations are possibly less willing to pay out large fees for consultant psychologists rather than training their own staff for the assessor position.

A major attraction of an AC is that it allows for managers or supervisors to be directly involved in the assessment process. This may help Human Resource practitioners to foster the commitment of managers to development and selection procedures, and therefore make a positive long-term contribution to Human Resource practices. Employing managers as assessors may also help to foster relations among staff members, especially trust in management. It may also signal the notion that the managers of a particular organisation are directly interested in the performance of staff members. Potential benefits could result from the practice of employing managerial assessors that could generalise to Human Resource systems across organisations. For example, the skills gained in AC assessor training could generalise to performance management systems.

Klimoski and Brickner (1987) suggested that one could assume psychologists would have less of an idea about the policy-capturing factors necessary for an individual to advance in an organisation when compared to line managers or supervisory assessors, who are highly cognisant of the bases of promotional decisions within a given organisation. This may also reflect the fact that managers are far more likely to hold tacit knowledge pertaining to the positions at hand, in terms of the intricacies and requirements that act as criteria for success in actual jobs. Given this, managers may be better equipped than psychologists to understand the criteria for effective performance in a given position. Such are the advantages associated with having an AC rated by a panel of job experts who have a comprehensive knowledge of the position under scrutiny (Paton & Jackson, 2002). A comprehensive source of direct job knowledge and experience will intuitively enhance the assessment quality by adding a rich source of expertise pertaining to the characteristics and/or behaviours that determine success in a given position. Such experiential knowledge is likely to be absent in a team of psychologist assessors.

Given that most organisations in the United States employ managers as assessors, it would appear that managers require a method and a paradigm by which to assess candidates that corresponds to their own schemata pertaining to successful AC performance. A method tailored for such a schema could theoretically yield the construct valid results that appear to be missing from many ACs that utilise managerial assessors. Moreover, the evidence suggests that the current trait-based paradigm in ACs is conceptually and practically problematic for managers. Treating a set of variables as trait measurements, when they are actually not rated as such, may lead to contamination of the ratings obtained. This contamination may extend into the means by which overall assessment ratings (OARs) are derived, and it is the OARs

that employment decisions are frequently based on in ACs. By treating AC ratings under a paradigm that does indeed fit in with the schema held by the managerial AC assessors, it is possible that gains in the predictive utility of the AC process may also be realised. Thus, the predictive utility of AC ratings obtained from managers may have been compromised in past research because they were applied under an inappropriate conceptual framework.

Attributing Variance to Both Exercise and Dimensional Features

The contention that variance in AC ratings should be attributed to both exercise and dimension features, was addressed thoroughly by Lievens and Conway (2001). Exercise features, in this context, refer to method variance or the correlation between different dimensions within a given exercise. Lievens and Conway found that the best fitting model for AC data across 34 studies reflected elements of both dimensions and exercises. This finding may have been attributable to the fact that 16 of the studies sampled stated that they primarily employed the services of psychologists as raters. Fourteen of the studies, on the other hand, employed managerial assessors, whilst 4 studies did not provide this information. In practice, a model of AC ratings that includes elements of exercises and dimensions could be problematic. Consider a developmental AC using such a model for feedback to candidates. Assessors attempting to give feedback on the basis of exercise and trait-based factors would firstly need to establish which factor, or combination of factors, contributed more variance to an individual's ratings. They would then need to explain to the candidate that feedback would be given on both elements because the measurement model did not fit neatly into either dimensions or exercises. This same scenario could logically generalise to selection decisions. The mixed

exercise/dimension model blurs the criteria under which selection decisions should be made. This presents a potentially awkward and unsatisfying situation for the practitioners who apply AC technology.

Overall Assessment Rating Integration Discussions in ACs

An issue of contention in the AC literature surrounds the common practice of having assessors discuss the information about each candidate at the end of the assessment process so as to derive overall assessment ratings (OARs) on the basis of these discussions (Ballantyne & Povah, 1995). Cook (1998) reports research findings that suggest the use of such concluding discussions at the end of the process may lessen the extent to which actual ratings are utilised optimally. Feltham (1989) found that weighting the four best predictors during the AC predicted criteria more effectively than the traditional OAR. Reilly et al. (1990) suggested that consensus discussions may be maladaptive, in that they are functionally time consuming, and research evidence suggests that they yield no potential predictive benefits. The authors suggested that using the total scores on behavioural checklists in an AC would lead to more optimal use of the assessment data.

The findings related to integration procedures to derive overall ratings are mixed, however, and this may be influenced by the lack of standardisation commonly found across users of the process (Fleenor, 1996). Most notably, Pynes and Bernardin (1992) found non-significant differences in the predictive validity of judgementally versus mechanically integrated ratings. In this study, the average predictive validity coefficients with a job performance criterion were .18 for the mechanical integration of ratings, and .15 for the integration discussion method. Pynes, Bernardin, Benton, and McEvoy (1988) further found that mechanical forms of integration and consensus

discussions lead to similar outcomes, except that the mechanical form resulted in substantial cost saving. Lowry (1988) also found these two forms of integration to be similar in their overall results when used for selection purposes. Lowry did, however, report changes in the rankings of individuals across the two methods of integration in a development centre. Further, Lebreton, Binning, and Hesson-McInnis (1998) found the judgemental integration of ratings to hold greater predictive qualities. The differences in overall magnitude between the two approaches, were however, found to be minimal.

The Measurement of Latent Constructs in ACs

Chan (1996) argued that since ACs that employ managerial or supervisory staff as assessors are predictive of relatively stable performance and promotability criteria, then they must be tapping into underlying constructs that are, likewise, relatively stable. It was suggested that the low construct validity often seen in the factor analyses of AC ratings may reflect the notion that ACs do not measure the intended constructs, but may demonstrate a form of latent‡ construct validity. Chan concluded that we may not be aware of what these latent constructs are as yet (Chan 1996; Spector, 2000). Russell and Domm (1995) presented the argument that although AC ratings do not tend to reflect the personal characteristics, skills, and abilities of AC participants and candidates, the AC process may still have construct validity. The suggestion was made that if ACs are able to capture criterion-related validity, then it follows that they are capturing some meaningful variance associated with relatively stable constructs. Exactly what those constructs are remains a mystery.

‡ The term 'latent' in this context differs from the terminology used in such methods as structural equation modeling (SEM) and item response theory (IRT) where multiple latent traits are investigated for the measurement of an underlying variable. The term 'latent traits', in the specific context of this section, refers to a variable that is measured by raters in a manner that is neither explicitly nor formally intended.

Klimoski and Brickner (1987) stressed the importance of understanding why ACs work, as no clear answer to this quandary is currently available. These sentiments still appear to be highly relevant in the present time. In consideration of the problem of construct validity associated with ACs, several suggestions were provided by Klimoski and Brickner, based on their review of the AC literature spanning ten years. The review attempted to offer explanations as to why ACs work, as their efficacy had been widely shown through predictive validity studies (Thornton, 1992). Alternative explanations were offered over the traditional trait paradigm by Klimoski and Brickner. The first explanations of AC construct validity covered actual criterion contamination and subtle criterion contamination. It was also suggested by these authors that latent constructs might be implicated in ACs and that these may, in turn, explain their predictive validity. These suggestions are manifest in the self-fulfilling prophecy/self-efficacy, and managerial intelligence explanations of AC construct validity. Chan (1996) suggested a further latent construct explanation in the form of impression management skill, and its role in determining success in ACs and managerial success. Finally, Klimoski and Brickner suggested that performance consistency might also explain AC construct validity. The ensuing discussion clarifies and elaborates on these concepts.

The Actual Criterion Contamination Explanation

The actual criterion contamination explanation (Klimoski & Brickner, 1987) suggests that the predictive relationship between assessment ratings and performance or promotion may be explained by the notion that only candidates who score well on an AC are considered for criterion decisions, such as promotion decisions, whereas others are not. The real issue of contention here is whether or not ACs are still

predictive, even when the criterion contamination is not present (Thornton, 1992). In other words, if those who are involved in promotion or performance decisions are not aware of how a given individual performed on an AC, are AC ratings still predictive of criterion measures of promotion and performance? Jones et al. (1991) rejected the actual criterion contamination explanation of predictive validity in their study, as the individuals involved in criterion decisions were not informed of OARs from the AC. Fleenor (1996) also discounted the possibility of actual criterion contamination explaining the predictive validity observed in a study of a developmental AC, by collecting criterion measures prior to the AC. A predictor/criterion relationship was still observed in this case.

In their meta-analysis, Gaugler et al. (1987) argued that actual criterion contamination was unrelated to the predictive validity of ACs. Regardless of whether feedback on ratings was given to candidates, whether the AC was used for academic research only, or whether the AC was used for decision making, comparable levels of predictive validity were obtained (see Table 2). Moreover, Thornton (1992) argued that since Gaugler et al. found that the accuracy in the prediction of various criteria was not a function of the type of criterion used, the actual criterion contamination explanation had insufficient foundation. In other words, ACs predicted such a myriad of criterion measures, for example performance, success in training, independent evaluation, etc, that it was unlikely that all would succumb to actual criterion contamination.

Table 2

Predictive Validity of Various Designs of AC Research

Design Approach	Estimated Predictive Validity
Experiment	.36
Predictive study: No feedback given to candidates	.43
Predictive study: With feedback given to candidates	.39
Concurrent Validity	.42

Source: Adapted from Thornton, (1992), with data from Gaugler et al. (1987).

The Subtle Criterion Contamination Explanation

OARs are often correlated with criteria related to promotion (Turnage & Muchinsky, 1984) and have even been found to correlate with promotion over and above performance-related criteria (Ballantyne & Povah, 1995; Chan, 1996). Such is the basis for the argument pertaining to the *subtle criterion contamination explanation* of AC criterion validity, first proposed by Klimoski and Strickland (1977) and again by Klimoski and Brickner (1987). This explanation suggests that instead of assessing candidates on the basis of particular stable underlying traits or personal characteristics, the assessors in ACs actually base their judgements upon factors that they perceive are important to 'get ahead' in a given organisation. Klimoski and Brickner suggested that rather than evaluating the specific abilities of an individual, assessors attempt to imitate the judgements of a promoter, and base their decisions on the policy-capturing factors of an individual. Such factors need not necessarily be related to actual performance on the AC, and therefore, AC ratings correlate with criteria related to promotion.

Mixed evidence exists for this explanation. The finding that ACs tend to predict promotion well, yet not necessarily performance, has been well founded

(Chan, 1996; Cohen et al., 1974; Hunter & Hunter, 1984; Schmidt, Ones & Hunter, 1992). Chan (1996) found that assessment ratings were predictive of promotion, but not supervisory ratings of performance, whereas, traditional cognitive tests were found to be more related to performance criterion measures, rather than promotability related criteria. Although a problem with ACs may be that results are contaminated in that they are used to help define future promotion decisions (Klimoski & Brickner, 1987), Chan's study was not contaminated in this way (Chan, 1996; Spector, 2000). Chan suggested that assessor ratings were perhaps tapping a 'policy capturing' construct (or cluster of constructs) that was distinct from those measured by traditional psychological tests. Such latent constructs must have also been distinct from those intended by the AC, as Chan's study found no evidence for construct validity. Turnage and Muchinsky (1984) similarly found that assessment ratings were unrelated to performance criteria, but were found to be predictive of promotion criteria. Turnage and Muchinsky reflected that these findings might add support to the subtle criterion contamination hypothesis.

Thornton (1992) argued against subtle criterion contamination with the contention that even with the multitude of criterion measures that have been used in research (Gaugler et al., 1987), ACs continue to sustain predictive validity (see Table 3). Thornton states that it would be difficult to maintain an argument suggesting that every one of these criteria were contaminated with judgement based on policy capturing factors or the image of what would seem to make a good employee in the eyes of assessors. Although contamination in this way may exist in some ratings based on performance or promotion criteria, some ratings are not predisposed to subtle contamination and have still been found to have predictive validity. Ballantyne

Table 3

Various Criteria Used and their Predictive Validity with AC Outcomes

Criterion Type	Estimated Validity
Performance	.36
Potential ratings	.53
Dimension Ratings	.33
Training performance	.35
Career progress	.36

Source: Adapted from Thornton, (1992).

and Povah (1995) argued that the differential ratings obtained from promotion and performance data may be an artefact of the criteria being assessed. They argue that promotion-related criteria often produce dichotomous data. In contrast, performance-related data require a focus on competencies and criteria, and thus have a greater potential for statistical error.

The Self-Fulfilling Prophecy/Self-Efficacy Explanation

The third rationalisation for the construct validity of ACs, proposed by Klimoski and Brickner (1987), was the self-fulfilling prophecy explanation. Under this explanation, being selected to participate in an AC and/or successful performance on an AC may act to reinforce feelings of self-efficacy within an individual. Theoretically, this leads the individual towards goal directed behaviour aimed at personal advancement within a company. This would possibly explain the relatively strong correlations between AC performance and criteria related to promotability.

The foundations of the self-fulfilling prophecy explanation find their roots in Bandura's (1982) notion of self-efficacy, and how this construct relates to performance in the AC context and criterion validation. Self-efficacy is defined as the judgement an individual passes on their own "capabilities to organise and execute courses of action required to attain designated types of performances" (Bandura, 1986, p.391). In the AC context, Klimoski and Brickner argue that the very action of being selected to partake in an AC may reinforce feelings of self-efficacy for candidates. The degree to which a person feels self-efficacious may govern the extent to which candidates exert effort and persistence (Bandura, 1986). Further, Klimoski and Brickner suggest that candidates who experience success during an AC may experience heightened levels of self-efficacy, and in turn may exert more effort and persistence towards goals associated with managerial success. As a result, these candidates may be more likely to gain promotion into managerial positions, where assessor's predictions will manifest through goal-directed behaviour.

Again, divergent evidence exists in regard to the notion of the self-fulfilling prophecy/self-efficacy explanation of AC construct validity. In their meta-analysis, Sadri and Robertson (1993) found a correlation of 0.34 between self-efficacy and work performance across 12 separate studies. Sadri and Robertson found that the correlation between self-efficacy and performance in simulated situations was higher (0.60). This may have implications for the strength of self-efficacy beliefs in the AC, as ACs are, in effect, collections of simulation exercises designed to simulate and represent relevant aspects of a particular job.

Jones et al. (1991) reported evidence against the self-fulfilling prophecy/self-efficacy explanation, as trainers and successful candidates of the AC used in their study were not informed of their OARs. However, the very notion of being successful

at all may have had an influence, even though in the Jones et al. study, individual candidates did not specifically know how well they performed compared with the other candidates. Gaugler et al. (1987) also found evidence against this explanation, as feedback given to assessors and candidates was not found to affect AC validity.

Evidence in favour of the self-fulfilling prophecy/self-efficacy explanation was found by Schmitt, Ford, and Stults (1986), who reported significant positive changes in self-perceptions as a result of participating in an AC. This change was found to occur when candidates were provided with specific feedback, and also in the absence of feedback. Fletcher and Kerslake (1992) investigated the effect of ACs and their outcomes on self-assessments. The authors also found that attending an AC had a positive impact on candidates' self-assessments, and that this endured over a long period of time. Furthermore, it was found that unsuccessful candidates had misjudged the effectiveness of their performance during the AC when compared to successful candidates. Thus, not all individual judgements may be accurate indicators of performance as judged by assessors.

This dissertation explores a particular aspect and extension of Klimoski and Brickner's (1987) original theme with respect to self-efficacy. Specifically it investigates the possibility that, in addition to playing a roll in AC performance, self-efficacy acts as a latent trait that can be detected by assessors. If manifest self-efficacy influences the decisions of assessors, then overall ratings of candidates demonstrating high levels of self-efficacy could be affected. Some evidence already exists for the other aspect of Klimoski and Brickner's explanation on the extent to which levels of self-efficacy are affected subsequent to performance on an AC (Fletcher & Kerslake, 1992; Schmitt, Ford & Stults, 1986).

The Managerial Intelligence Explanation

Because ACs tend to predict criteria related to promotion, the possibility exists that the candidates who perform well in ACs hold a knowledge of what it takes to get ahead in managerial contexts. That is to say, they are savvy about, sensitive towards, and skilled in, situations that require managerial knowledge. Along the lines of these arguments, Klimoski and Brickner (1987) suggested a *managerial intelligence explanation* that sought to explain the construct and criterion-related validity of ACs. This theory, which is related to the subtle criterion contamination explanation, stems from the notion of tacit knowledge (Wagner & Sternberg, 1985). Tacit knowledge is referred to as knowledge that is not openly expressed, or which remains latent. It incorporates learning from experience, and reflects the ability to use existing knowledge flexibly, or to adapt existing knowledge to understand and comprehend new experiences. Tacit knowledge can be thought of as an appropriate response to one's natural environment. Moreover, it involves an adaptive response to 'real world' situations as opposed to academic situations. The construct involves problem solving to facilitate the well-being, goals, needs, and survival of an individual (Wagner & Sternberg, 1986).

Tacit knowledge is characterised as the knowledge that is gained from everyday experience that is both implicit and unarticulated (Sternberg, Forsythe, Hedlund, Horvath, Wagner, Williams, Snook & Grigorenko, 2000). It is the practical ability that allows people to learn from the experiences they have, and to apply the knowledge that they have learned in a goal directed manner (Wagner & Sternberg, 1985). Tacit knowledge is useful for adapting to and shaping real-world environments, and is therefore an important consideration in the successful performance of practical tasks in a multitude of domains. Sternberg et al. (2000)

suggest that the acquisition of tacit knowledge involves four major components. Firstly, it is acquired with little or no environmental support from media or others that help people to acquire it. As a result, tacit knowledge usually remains latent and under-emphasised in terms of its relative importance for performing tasks. ACs may provide an arena in which individuals are apt to obtain tacit knowledge through its associated mechanisms. For instance, this might include the assessee's emphasis on selective encoding (gleaning more information from less), selective comparison (recall of symbolic memories that are relevant to the current practical situation) and selective combination (i.e., combining information in goal directed ways) (Blanchard & Thacker, 1998; Sternberg, et al., 2000).

Secondly, there exists the notion that tacit knowledge is inherently procedural, in that it is associated with practical action. It guides behaviour, very often in an automatic or unconscious manner. Tacit knowledge tends to be constructed of a complex combination of specified goal directed, and often multi-conditional, procedural rules. Thirdly, tacit knowledge is useful in the practical environment. It is instrumental in assisting an individual to attain a specified goal. Lastly, tacit knowledge has coherent relations among its features. The three components which convey that tacit knowledge is often acquired on one's own, is procedural, and is instrumental in attaining goals, all fit together meaningfully. Procedural knowledge is often practically useful because it is concerned with how to perform a given task. Knowledge acquired largely by the individual alone is more likely to be practically valuable, in terms of holding relevance to real-world situations that the person has or will encounter. As procedural knowledge is often difficult to verbalise, it is more likely to be gained experientially.

Sternberg et al. (2000) suggested that AC exercises provide a useful context in which tacit knowledge can be manifest, with the advantage of closely representing performance in a given domain. Klimoski and Brickner (1987) have suggested that tacit knowledge may present a major component of what ACs truly measure, rather than the ability traits that traditional ACs purport to measure. Tacit knowledge has been related to performance across many different domains and managerial and occupational groups. These include corporate organisations, military leaders, sales people, banking staff, as well as cross-cultural applications (Sternberg et al.). There is strong evidence to suggest that tacit knowledge may well be a strong determinant of performance. The possibility that ACs may allow for a manifestation of this elusive construct possibly accounts for the strong predictive qualities of the AC paradigm.

Wagner and Sternberg (1986) refer to a three-category framework under which the theory of tacit knowledge can be understood; comprising tacit knowledge about managing one's self, managing others, and managing career. Tacit knowledge about managing self involves managing one's activities in such a way that productivity is maximised. This concerns the prioritisation of tasks, maximising effectiveness in terms of output of effort and self-motivation. Tacit knowledge about managing others concerns the management of organisational members and social relationships. Tacit knowledge about managing one's career involves knowledge pertaining to the establishment of careers, how reputations are refined and the persuasiveness of the individual.

Colonia-Willner (1998) found that the Tacit Knowledge Inventory for Managers (TKIM) (Wagner & Sternberg, 1991) predicted managerial skill in a sample of bank managers, whereas more traditional measures of academic intelligence did not. It was reported, as a limitation to the study, that the TKIM predicted

performance poorly for a large part of the sample, yet the TKIM predicted factors such as income much better. This finding may be related to the propositions of Turnage and Muchinsky (1984) who found that the AC in their study better predicted criteria related to advancement, as opposed to performance. Thus, in some circumstances, tacit knowledge may predict criteria related to promotability or advancing one's position within an organisation.

Research evidence suggests that estimates of IQ and tacit knowledge measure two quite distinct constructs (Sternberg, et al., 2000). Klimoski and Brickner (1987) argue that measures of IQ consistently and positively relate to performance on the job, however, these relationships tend to be moderate (Cook, 1998). Klimoski and Brickner also cite several articles that have found a relationship between traditional measures of IQ, and OARs in the AC context. Scholz and Schuler (1993) found a corrected correlation of 0.43 between intelligence scores and OARs. Thornton (1992) asserted that the real issue of interest concerning the managerial intelligence explanation in ACs is whether ACs actually do measure something useful, over and above IQ tests.

Sternberg et al. (2000) reported non-significant correlations between IQ and tacit knowledge across a number of domains, including undergraduate samples, business executives, air force recruits and a sample from a rural village in Kenya. Wagner and Sternberg (1991) reported evidence that the TKIM predicted performance-related criteria more effectively, and independently of measures of IQ and other traditional psychological tests. Likewise, Wagner and Sternberg (1990) reported evidence that tacit knowledge had stronger predictive validity than cognitive tests, personality inventories and interpersonal orientation. Despite the strong possibility that tacit knowledge may well play a role in AC technology, no literature

to date could be found that has directly investigated Klimoski and Brickners' (1987) suggestion that tacit knowledge may have a positive relationship with OARs and criterion measures used in ACs.

The Impression Management Skill Explanation

Chan (1996) suggested a further dimension that may explain the relationship between OARs and criteria related to advancement or promotion. The *impression management skill explanation* presents the argument that individual differences exist in terms of the skill and efficiency by which people employ different impression management strategies and techniques. In ACs, individuals who are skilled at identifying appropriate behaviours that will elicit favourable impressions from assessors may be more able to create favourable impressions with supervisors on the job, thereby leading to higher levels of promotability. Chan argues that in ACs, the awareness of being evaluated may evoke self-presentational concerns, and thereby make the elicitation of impression management strategies more prominent.

Tedeschi and Riess (1981) suggested that impression management behaviours consist of any behaviour intended by a person to control or manipulate the impression they express to others. The most important implication of this behaviour is not how the actor views his or her own behaviour, but rather, the attributions made by the observer (or observers) on the basis of observed behaviour. Such self-presentational strategies can be used in order to gain immediate objectives (Tedeschi & Riess, 1981). An individual's expectancies can also influence the onset of impression management behaviours, in terms of their own judgement of the probability of their success on a given task (Arkin, 1981).

Baumeister (1982) suggests that an individual may be driven to perform at a given level on the basis of self-presentational concerns, so as to please an audience or to seek social approval. One important form of impression management, which may be implicated in such self-presentational concerns, is that of self-monitoring behaviour (Kolb, 1998; Snyder, 1974). Snyder (1974) distinguishes between high self-monitors as individuals who are able to control the image that they portray in social interactions, and low self-monitors as individuals who strive to keep congruence in regard to who they are and how they behave (Snyder, 1987). Snyder suggests that individual differences exist in self-monitoring behaviours in that some people are more sensitive than others about the impression they convey in social situations.

The measurement and psychometric conceptualisation of self-monitoring behaviour has become somewhat of an issue of contention between various researchers (Snyder, 1987). Originally, Snyder (1974) developed a self-monitoring scale which conceptualised the construct as including items which sought to assess five components associated with the construct: (a) concern for the social appropriateness of one's self-presentation; (b) the use of social comparison information as cues for appropriate self-expression; (c) control and modification over one's self-presentational and self-expressive behaviour; (d) use of this ability in particular situations; and (e) the extent to which expressive and self-presentational behaviour is cross-situationally consistent or variable (Snyder, 1974, p. 529).

The Snyder (1974) scale was the only measure of the self-monitoring construct available, until a notable critical evaluation of the psychometric properties of the scale was conveyed by Lennox and Wolfe (1984). Lennox and Wolfe reported that several studies (e.g., Briggs, Cheek & Buss, 1980; Gabrenya & Arkin, 1980) had found that

the 5 components proposed by Snyder were not being measured, and that only three factors were appearing on factor analyses of the Snyder scale. These factors were: (a) acting ability; (b) extraversion, (c) other-directedness. As the Snyder scale did not appear to measure the constructs as theorised, Lennox and Wolfe reported that the internal consistency for the Snyder scale is generally found to be middling, with Cronbach alpha coefficients not exceeding .70.

Furthermore, Lennox and Wolfe (1984) reported that the total score on the Snyder (1974) scale was not interpretable, as the various factors within the construct were in competition with each other. Consequently, Lennox and Wolfe constructed an alternative scale. Based on Snyder's (1974) review of self-monitoring, Lennox and Wolfe developed a narrower conceptualisation of the construct with a 13-item measure of just two components thought to be important in self-monitoring: (a) sensitivity to the expressive behaviour of others; and (b) ability to modify self-presentation (Lennox and Wolfe, 1984, p. 1361). The revised scale yielded promising psychometric integrity in its original study, with internal consistencies exceeding .70, with two sub-scale scores that combined to form a total score.

In a rebuttal, Snyder (1987) argued that that his original scale and the Lennox and Wolfe (1984) scale correlated at .72 when corrected for attenuation, thereby forming an argument against the notion that Lennox and Wolfe had, in point of fact, developed an entirely new measure. Snyder argued that the Lennox and Wolfe scale was too narrowly focussed, with items tending to re-state each other too often, and, with only two items reversed scored, was at risk of response bias. Items were criticised as being overly lengthy, ambiguous or using unusual language. However, measurement choice decisions made by subsequent researchers (see below) suggest

that avoiding the factor structure problems in the Snyder scale was viewed as more important than concerns about the counter issues against the Lennox and Wolfe scale.

Much of the recent research has tended to opt for the measurement of self-monitoring through one of three approaches: 1) the Lennox and Wolfe (1984) scale (Anderson, Silvester, Cunningham-Snell & Haddleton, 1999; Kolb, 1998), or 2) a combination of the 'best' items from Lennox and Wolfe (1984), Snyder (1974) and from Leary's Motivation to Impression Manage scale (1990, as cited in Warech, Smither, Reilly, Millsap & Reilly, 1998), or 3) the use of the Lennox and Wolfe scale, with a slight modification in the wording of items aimed at lessening ambiguity (O'Cass, 2000). These studies have generally reported making psychometrically and factorially sound measurements of the self-monitoring construct, largely under the influence of Lennox and Wolfe.

Empirical evidence suggests that impression management behaviour may have a bearing on the performance of tasks in the workplace (Baumeister, 1982). Snyder (1987) reports positive correlations between high self-monitoring individuals and job performance in positions requiring high levels of interaction and communication. These positive correlations were also reported with job level, in that high self-monitors tended to be managers, with low self-monitors tending to hold technical or clerical positions. Whitmore and Klimoski (1984) found that high self-monitors tended to become the leaders in groups involving problem-solving tasks, which perhaps reinforces the importance of the construct in the AC context.

High scores on AC ratings may be influenced by the degree to which an individual holds a propensity to take a leadership role in group situations. Research has suggested that high self-monitors may hold some of these inclinations. Kolb (1998) found small positive correlations between self-monitoring and leader

emergence in student groups. Students who were high self-monitors tended to be reported more commonly as leaders by their peers. Snyder suggested that this tendency for high self-monitors to become leaders might be due to their inclination to initiate conversations. This may generalise, in the group context, to facilitating more rewarding interactions with group members, thereby positioning high self-monitors as leaders. Kolb's findings were consistent with that of previous research, which suggests that high self-monitors tend to emerge as leaders more often than low self-monitors (Day, Schleicher & Unckless, 1996).

The self-monitoring construct may hold even more importance in ACs, as further research suggests that high self-monitors employ certain tactics to become leaders in groups. Snyder (1987) conveyed further empirical findings, regarding the relationship between self-monitoring behaviour and leadership, by reporting some of the tactics that high self-monitors employ in situations requiring leadership. These strategies included goal setting, supportiveness, and motivating and encouraging others. Snyder emphasised that high self-monitors tended to become leaders in situations requiring high levels of verbal interaction. Further to this, Felson (1981) suggested that high self-monitors, who are by definition concerned about their own behavioural appropriateness both interpersonally and situationally, would be particularly sensitive to the presence of an audience. This notion ties in with Chan's (1996) suggestion that ACs provide an audience composed of assessors, and that this notion may heighten the salience of self-presentational concerns to the assessee.

Warech et al. (1998) studied the relationships between self-monitoring, and 360 degree ratings for managers participating in a developmental AC. The study found that self-monitoring ability was associated with job-related interpersonal effectiveness (operationalised as empowerment, managing teams and influencing

others), suggesting that the managers in the study who had greater self-monitoring skills were able to respond effectively to situations requiring interpersonal skills or job-related interpersonal effectiveness. Self-monitoring ability was, however, unrelated to peer ratings of business competence. This finding may provide some further clues as to why AC ratings often correlate with such criteria as promotion, but not necessarily with 'on the job' performance (Ballantyne & Povah, 1995; Chan, 1996; Turnage & Muchinsky, 1984). In line with Chan's (1996) suggestions, it may be that self-monitoring plays a role in the extent to which individuals can get ahead or 'policy-capture' in an AC. These possible policy-capturing skills held by self-monitors may extrapolate to criterion measures related to promotion, but not necessarily actual job performance. Moser, Diemand and Shuler (1996) looked at AC overall ratings and two components of self-monitoring behaviour: the level of inconsistency in behaviour and the extent to which individuals displayed social skills. No relationship was found between the inconsistency component of self-monitoring and outcomes on the AC, although a small correlation was reported ($r = .26$) between the social skills component of self-monitoring and AC ratings.

Intelligence, Personality, and their Relationships with OARs

Researchers have investigated relationships between several other psychological variables and assessment centre overall scores. Indeed, most of the psychological variables that have been investigated have been measured externally to the ratings obtained in the AC or as an adjunct to the AC ratings. Overall, these studies suggest that a moderate to small amount of the variance in OARs can be explained by intelligence and personality based measures. Scholz and Schuler (1993) investigated the relationship between some AC relevant constructs and the constructs'

relations with AC ratings. The authors found, via meta analysis methodology, that self report measures of intelligence explained 18% of the variance in OARs. Likewise, external measures explained the following amounts of variance in OARs: interpersonal competence (9.60%), achievement motivation (9.00%), dominance (5.29%) and self-confidence (6.76%) Curiously, it was found that agreeableness was not related to AC outcomes. The authors concluded that the overall ratings in ACs were influenced to some degree by certain traits associated with career advancement that were not explicitly measured.

Schmidt and Hunter (1998) found that general mental ability explained 25% of the variance in OARs. Fleenor (1996) found small correlations between AC outcomes and the trait 'exhibitionism', which is characterised by individuals who have a preference for being the centre of attention, enjoy having an audience, and enjoy being dramatic and witty. In personality theory, one study found that AC ratings correlated with conscientiousness and extraversion on the NEO (Furnham, Crump & Whelan, 1997). A study by Goffin, Rothstein, and Johnston (1996), however, reported no clear relationships between AC ratings and personality measures.

Some of the constructs found to be related to AC outcomes theoretically have convergent relationships with some of the constructs suggested by Klimoski and Brickner (1987) and Chan (1996). The relationship between intelligence scores and performance on ACs has been demonstrated in other studies (Schmidt & Hunter, 1998). Impression management skills, self-monitoring social skills and interpersonal competence with the possible addition of self-confidence, extraversion and dominance, perhaps a similar manifestation to social confidence, and possibly exhibitionism, appear theoretically related (Chan, 1996; Fleenor, 1996). Self-efficacy

may also be related to achievement motivation and high self-esteem (Schmidt & Hunter, 1998; Scholz & Schuler, 1993).

The Behavioural and Interactionist Paradigms

In this section, the behavioural and interactionist paradigms are discussed as a background to other explanations of AC construct validity that contrast with the traditional trait paradigm. Such approaches offer alternative explanations of how ACs predict performance and promotion in the work place. One such approach is the applied behavioural analysis paradigm, summarised by Delprato and Midgley (1992). Under behaviourism, behaviour is understood to operate under a lawful basis, and is understood to be determined largely by environmental factors. Skinner (1974) suggested that under this model, a scientific understanding of the determinants of behaviour should allow for the prediction and control of behaviour.

A further fundamental notion, under the behavioural model, is the denial of the notion of dualism (Skinner, 1974). Skinner argued that only physical or material events exist in the world. Moreover, the behavioural approach asserts that the study of psychology should only be concerned with the study of behaviour – i.e., that which is directly accessible and measurable (Delprato & Midgley, 1992). This contention argues against mentalistic explanations that attempt to move beyond the realm of human behaviour, and seek to infer the existence of mental entities such as traits. This approach asserts that the causes of behaviour are found in the environment, and that organisms change through altering contingencies of reinforcement in their environment (Skinner, 1974).

Contemporary approaches to behaviourism have added a cognitive emphasis, particularly with the work of Bandura (1977) and Mischel (1973). Mischel (1968) made a vital, albeit controversial suggestion that has particular significance in terms of offering an alternative explanation for the original exercise effect findings of Sackett and Dreher (1982). Currently, practitioners and researchers of ACs adhere to the trait paradigm when rating candidates (Lowry, 1996; Sackett & Harris, 1988; Spychalski, et al., 1997). This stance assumes trans-situational consistency in behaviour across AC exercises. In his original monograph, Mischel (1968) argued that individuals showed much less consistency in their responses cross-situationally than purported by trait theories, and that this lack of cross-situational consistency holds across both highly similar and highly differentiated situations (Mischel, 1968, p. 177). Mischel's arguments challenged the strongly held notion that traits are relatively stable and highly situationally consistent attributes with generalised causal effects on human behaviour (Mischel, 1973). Although some authors have argued that the dimensions assessed in ACs should not be treated as trait-based variables (Sackett, 1987) but rather as categories for behavioural items (Byham, 1980), a reasonable take on the literature reviewed would suggest that quite strict trait-based models prevail in AC methodology, which is manifest in both research and practice.

Although Mischel's work is most commonly associated with personality attributes, the term trait can refer to "any distinguishable, relatively enduring way in which one individual varies from others" (Guilford, 1959, p. 6). This definition encapsulates personality constructs, underlying characteristics, qualities, or processes, inferred by measuring behavioural indicators (Mischel, 1968). As a result, the factors that Mischel identified regarding situational effects on behaviour and/or behavioural

measurement are likely to be highly relevant to the traits that are intended for measurement in ACs.

Mischel's research added empirical evidence to his theories, with findings suggesting that an individual's behaviour is not necessarily cross-situationally consistent (summarised in Mischel, 1984). Mischel also found evidence that individual differences, situations, and response modes, account for more variance when sampled together than alone. Subsequently, an uproar ensued in the field of psychology, for which some of the basic elements of research, particularly concerning personality, had been challenged (Kenrick & Funder, 1991). Mischel's original arguments, however, had been misconstrued by several theorists as taking an anti-dispositional stance (Mischel, 1973). Mischel's arguments actually maintained that an individual's behaviour in any situation may be changed distinctly by minor situational alterations. He questioned the utility of only inferring generalised dispositional characteristics from behavioural signals alone as a foundation for understanding an individual's behaviour (Mischel, 1968; 1973; 1984).

The acknowledgement of the influence of the situation on behaviour called for an alternative means by which to investigate the discriminatory and consistent characteristics found in human social behaviour (Mischel, 1973). As an alternative to inferring the existence of broad, generalised dispositional characteristics, Mischel suggested that a more useful method would be to observe how people actually behave, relative to the situation. This would be a move toward an investigation into "direct behavioural samples and reports relevant to the particular problem, outcome, or domain of interest and anchored to the specific social and psychological context" (Mischel, 1984, pp. 352-353). For instance, Sackett and Dreher's (1982) original factor analyses showing the lack of evidence for convergent and discriminant validity

in ACs could be interpreted as error, or could be reinterpreted to show that by linking behavioural measures to specific contexts, ACs largely act as a group of divergent contexts, the exercises in ACs act as though they are direct behavioural samples relevant to a particular situation or context.

Taken together, the contemporary stance on these perspectives is that human behaviour is best conceptualised neither by the situation nor by dispositional characteristics alone. Rather, in order to understand a person's reactions holistically, both situational and dispositional characteristics need to be acknowledged. This view has been labelled the *Interactionist Paradigm* (Bem & Funder, 1978; Tett & Guterman, 2000) and acknowledges that seemingly trivial changes in an individual's environment can influence behaviour (Highhouse & Harris, 1993). The next section discusses how certain aspects of the behavioural and interactionist paradigms might be useful in terms of creating a basis for AC ratings in practice.

The Performance Consistency Explanation

The behavioural and interactionist line of thought has contributed to a major component of the performance consistency explanation of AC ratings (Klimoski & Brickner, 1987) by offering an alternative to the prevailing model of assessment in ACs. The performance consistency explanation presents a comparatively radical perspective on AC ratings, as it openly questions the trait-based foundations of the process from two different angles. While the first part of the explanation is a critique that research would suggest is not particularly cogent, the second part of the explanation presents a view that may actually help to explain the reality behind AC ratings.

The first part of this argument presents the possibility that the stable traits being judged across exercises may be erroneous, and that a different process may be at work. Klimoski and Brickner suggested that ability traits may not be used by assessors at all in ACs, and that performance may be judged by evaluating past and present performance of individuals on the job and basing judgements on these factors. According to Klimoski and Brickner, assessors could be made aware of past and present performance through biographical information presented to them about each assessee. This aspect of the performance consistency explanation is notably weakened by the notion that many ACs keep the bio-data and identity of assessees confidential from assessors, and yet the ACs remain predictive (Thornton, 1992). Furthermore, Turnage and Muchinsky (1984) found little support for the contention that their bio-data component accounted for the predictive validity of their AC.

The second part of the performance consistency explanation centres on the idea that the exercises contained within ACs act as work samples of behaviour. AC exercises are preferably constructed from job analysis data (Ballantyne & Povah, 1995), and it is generally inferred that if candidates perform well on exercises that they have the necessary knowledge, skills, and ability traits to perform a particular job effectively. Several authors argue that there is an unnecessary insistence in psychology that predictors of performance need to be different from the criterion that they predict. This may not necessarily need to be the case, as the predictive validities of work samples show (Campion, 1972; Muchinsky, 2000; Robertson & Kandola, 1982; Robertson, et al., 1987; Schmidt & Hunter, 1998; Schmitt & Ostroff, 1986). The work sample approach is explained theoretically under *Behavioural Consistency Theory*, which asserts that past behaviour is the most accurate predictor of future behaviour, and that like behaviour is predictive of like behaviour (Cook, 1998). Note

that the emphasis under the behavioural consistency theory is behaviour and behaviour only. These notions generally belong to the behaviourist school of thought, under which the inference of higher-level traits is never made.

The behavioural consistency approach helps to ensure high levels of predictive validity, as there is a close correspondence between that which is being predicted (criterion measures) and the predictors themselves (Cook, 1998). Given that ACs are composed of multiple simulations, ACs that are designed for behavioural assessment through particular simulation exercises, as opposed to the measurement of stable traits, could thus be advantageous. Such advantages could include improved predictive validity and improvements in the quality of developmental feedback. Such an assessment may potentially avoid issues pertaining to the construct validity of trait-based variables. Sackett and Dreher (1982) suggested that low monotrait-heteromethod correlations could be due to situationally determined behaviour. If the ratings in ACs consistently emerge as exercise ratings rather than dimensional ratings, and ACs still remain predictive of performance and potential, then perhaps psychologists should design their ACs to focus on task-specific measurement. The inference of higher-level traits may not be appropriate in a traditional AC format. Thus, AC ratings may show the extent to which an individual performed well on individual exercises. Carrick and Williams (1999) suggested further that the exercise effect makes it difficult to draw conclusions pertaining to the level of an individual's skill that may transfer across divergent situations. Gorham (1978), Lowry (1995), and Robertson et al. (1987) have suggested that the AC paradigm should be designed so as to treat the exercises as stand-alone work samples, rather than to make the inference of stable ability traits. Under this model, the behaviour exhibited by an individual would be assessed by means of exercise specific behaviourally-based rating scales.

A clear example of the exercise effect in ACs was presented by Robertson et al. (1987). These authors factor analysed the assessment data they obtained from three ACs using non-psychologist assessors, and found the resulting factors to be clearly interpretable in terms of exercise dimensions, rather than trait dimensions across all three centres. Robertson et al. suggested that despite the fact that common terms existed across exercises for a given dimension, assessors may have been giving different interpretations to the labels across exercises. In practice, Robertson et al. suggested using the original job analysis data, and ascribing assessment dimensions based on these data, without inferring underlying traits that may account for these tasks or behaviours. In terms of calculating OARs under this alternative approach, Lowry (1997) and Robertson et al. (1987) suggested that a useful substitute for measuring ability traits in AC exercises could be derived from OARs that represent the sum total performance on individual work simulations. In general, situationally specific exercises have a history of displaying strong predictive validity coefficients in the work sample literature (Campion, 1972; Hunter & Hunter, 1984; Robertson & Kandola, 1982; Schmitt & Ostroff, 1986).

When considering the use of ACs that form their bases from work sample methodology, attention must be drawn to the validity of the work sample as a predictor of behaviour. Asher and Sciarrino (1974) reviewed the literature relating to the predictive validity of high-fidelity (highly work related) and low-fidelity (work related, but to a lesser extent than the former) work samples. Across 42 high-fidelity work sample studies, 43% had predictive coefficients of no less than .50 with job proficiency as the criterion. Seventy per cent of these studies had predictive coefficients of not less than .40 with job proficiency. Lower-fidelity work samples appeared to relate better than high-fidelity samples to training criteria. Thirty nine per

cent of low-fidelity work samples related to training criteria with a coefficient of no less than .50, and 65% no less than .40. Hunter and Hunter (1984) re-summarised this literature, and reported mean corrected predictive validity coefficients for verbal work samples with a training criterion (.55) and for a proficiency criterion (.45). For motor work samples, the validity coefficients were also reported for the training criterion (.45) and for the proficiency criterion (.62).

In their comprehensive meta-analytic work across thousands of validity studies, Hunter and Hunter (1984) found a corrected mean correlation between work samples and job performance of .54. These data were summarised again in Schmidt and Hunter (1998), and indeed work samples were found, in this study, to be the strongest singular predictor of a performance criterion of the 19 forms of prediction surveyed. With a narrower pool of validity coefficients to select from, Schmitt, et al. (1984) found modest validity coefficients for work samples and various criterion measures. Across studies published in the Journal of Applied Psychology and Personnel Psychology between the years of 1964 and 1982, mean correlations were reported between work samples and various criteria, such as performance (.32), achievement/grades (.31), wages (.44) and with other work samples (.35). Overall validity was reported as .38.

Lowry (1995) asserts that the 1989 international AC guidelines (Task Force on Assessment Center Standards, 1989) allow assessments to be made in terms of tasks, as opposed to only dimensional characteristics, allowing a broader choice in terms of how behaviour is assessed. The current guidelines (International Task Force on Assessment Center Guidelines, 2000) however, refer to the use of the more general term 'competencies', which can nonetheless be extended to encompass tasks, due to

the somewhat nebulous definition surrounding the term 'competency' (Woodruffe, 1993)

Several pioneering suggestions on task-based ACs have arisen from one particular researcher by the name of Lowry. Lowry (1995; 1997) introduced terminology, which assists in summarising and crystallising the arguments presented earlier, and furthermore makes the distinction between dimension-specific and task-specific ACs. A dimension-specific process refers to any AC that attempts to assess stable higher-level ability traits such as leadership or communication skills (i.e., the traditional trait model of AC technology). In contrast, in a task-specific AC, the behavioural responses associated with a specific AC exercise are assessed, as opposed to making inferences as to internal mechanisms underlying behaviour (Lowry, 1995). Lowry's work is based on the same premise as the other suggestions made in this section; that predictive validity coefficients are a function of the extent of similarity between predictor variables and criterion variables. As Robertson and Kandola (1982) suggest, the ultimate goal is to achieve the highest predictive validity coefficients possible. The power of such correlations may be weakened when predictors and criteria are distinct from one another. In support of this view, Russell and Domm (1995) found that task-based ratings lead to greater construct validity than the traditional ability trait ratings.

A task-specific approach also lends itself toward a type of realistic job preview for candidates, in a similar mode to work sample tests (Herriot, 1986). There is research evidence to suggest that the use of work samples can reduce turnover rates in this way (Cascio & Phillips, 1979). Herriot also suggested that a task-specific approach would make task-based training needs more easily identifiable, as well as

making the assessment process procedurally more simplistic, as tasks could be subsequently used as job performance indices.

The simplicity of this approach avoids some complex questions that may arise when inferring traits from behavioural data. For instance, such complex notions as frame of reference training could be simplified under a task-specific approach. This is because there is no need for assessors to hold a shared understanding as to what constitutes behavioural manifestations of particular trait categories. A shared understanding of varying levels of behavioural performance would almost certainly be beneficial, without the complications of having to categorise that behaviour into a trait-based framework. Forming such a frame of reference in a task-specific approach would be straightforward, as explicit behavioural indicators are already provided to assessors in the form of a checklist. The assessor does not seek further information nor is further classification necessary beyond behavioural ratings.

Herriot (1986) suggests that trait judgements may involve issues of ambiguity surrounding which categories of tasks require the existence of certain specified attributes. Ambiguous inferences with regard to certain behaviours being *manifestations of underlying traits* may be problematic or may involve an awkward degree of subjectivity. Questions surround whether one can generalise from the results obtained in one situation to another. Problems may arise when the intention is that assessors should hold a shared mental model in terms of which behaviours exemplify which traits and whether these traits are actually independent of one another. Assessors could also be potentially distracted by attempting to infer attributes that are not relevant to the current assessment.

Lowry (1995, p. 444) summarised the advantages of using task-specific ACs from his own research as including:

1. Greater approval for the process from assessors in terms of the ease by which behaviour can be assessed.
2. Increased inter-rater reliabilities (reported as exceeding .80).
3. Task-specific training ensures complete familiarity with exercises.
4. Behavioural checklists can be constructed so as to maximise clarity for assessors in terms of what constitutes acceptable performance on a given exercise.
5. Behavioural feedback was found to be more advantageous than dimensional feedback in terms of forming a useful understanding for centre participants.

Evidence In Favour Of A Task-Specific Approach

Thornton, Kaman, Layer and Larsh (1995) presented evidence for the efficacy of the task-specific approach in a study that found feedback given to AC participants based on their exercise performance resulted in greater behavioural change than did trait feedback. In addition, Adams (1990) found that ratings associated with performance on tasks, as opposed to dimensions, increased the extent to which participants were able to recall their specific behaviour on the AC exercises. Moreover, other research has found that feedback based on traits such as intelligence has been associated with negative consequences on behaviour in samples of children (Mueller & Dweck, 1998). Particularly, this research has found that trait-based feedback was associated with less task persistence, less task enjoyment, lower ability attributions and lower task performance than individuals who received feedback based on effort.

Further evidence for this approach was found by Lance, Newbolt, Gatewood, Foster, French, and Smith (2000). The series of studies presented by Lance et al. correlated exercise factors, that is correlations between different traits within a given exercise, with externally measured constructs such as cognitive ability measures. Lance et al. found some evidence of expected relationships between exercise factors and externally measured correlates, and concluded in general that exercise effects reflect true sources of participant variance, rather than halo effects. Lance et al. argued that AC ratings in their study were best explained by a global person related factor, that is the overall performance of an individual on an AC, with the addition of exercise variance.

Carrick and Williams (1999) suggested that ACs which derive their ratings on the basis of performance on individual tasks or exercises may be advantageous in settings where the purpose is selection, recruitment and promotion (which are usually the main aims of ACs). Development centres focus on assessing an individual's strengths and weaknesses, whilst developing a plan for future training needs (Cook, 1998). In development centres, Carrick and Williams suggest that the task-specific AC may become problematic when attempting to diagnose and develop personal strengths and developmental needs. This may, however, not be the case, given the previously discussed findings suggesting the advantages of giving task-specific feedback in terms of fostering greater behavioural change and aiding recall of specific learned behaviours (Adams, 1990; Mueller & Dweck, 1998; Thornton et al., 1995).

Another argument against the use of task-specific ACs could stem from the notion that by using work sample type exercises in the measurement of candidate abilities, all of the limitations associated with work samples are inevitably encountered. The major relevant limitation with samples of behaviour is the

suggestion that work samples measure performance at a given time, but do not assess potential performance of the individual (Muchinsky, 2000). This is hypothetically problematic, as ACs are celebrated for being able to predict future criteria (Carrick & Williams, 1999). Lane (1992), however, argued that if an AC is designed such that it assesses a sample of a future job, or aspects of a future job, then the assessment may well be predictive of future performance. Robertson et al. (1987) suggested the use of task analyses when constructing task-specific ACs are crucial for identifying the key tasks that are important in managerial roles. Such fundamental tasks would be likely to last over time, and would most likely be beneficial in predicting future job performance. Theoretical arguments against work samples being predictive of future performance seem dampened by their relatively strong predictive coefficients reported in research (Hunter & Hunter, 1984).

Neidig and Neidig (1984) and Lance et al. (2000) noted that the exercise effect so pervasively seen in AC research might reflect a 'real' exercise effect. It may be that individuals perform better on some exercises as opposed to others due to the fact that they hold more of an aptitude for the content of some exercises over others. From a task-specific perspective, ACs may indeed hold construct validity, in terms of exercise factors. The previously mentioned studies have shown clear factor loadings on exercise dimensions. It may be that the work sample is the best operational paradigm for ACs, while retaining the process as it stands without forcing it into the trait paradigm.

While some studies (Highhouse & Harris, 1993; Lance et al., 2000; Neidig & Neidig, 1984) have reported that exercises themselves may reflect true and useful variance, there is a potential problem associated with this notion. Exercise effects, according to these studies, are comprised of strong correlations between different

dimensions measured within a given exercise. There is evidence to suggest that the prevailing population of assessors are not particularly skilled at making trait-based judgements, even with small numbers of traits (Gaugler & Thornton, 1989). Indeed, as previously discussed, there are multiple sources of evidence to suggest that exercise effects are strong, at least among managerial assessors. The deliberation surrounding this evidence questions whether managerial assessors really make trait judgements at all. Lance et al. (2000) suggest that the trait-based evaluation process in ACs is too complex to be carried out as intended. A less complex approach would be to have assessors rate situationally specific behaviours and treat assessment exercises as stand-alone work samples of behaviour, whilst avoiding trait-based judgements altogether. Such a task-specific approach would hold the advantageous characteristics associated with making greater conceptual sense, increased measurement precision, ease of training and more productive behaviourally-based feedback.

One might question how a behavioural approach, such as this, would explain some of the consistencies in behaviour that have been reported in the literature (Lance et al., 2000). Although the situation is said to hold a profound influence over behaviour (Mischel, 1984), behaviourism offers explanations for situational behavioural consistency in terms of behavioural repertoires. In this view, people tend to show some level of behavioural consistency due to the fact that they have acquired some stable response tendencies through experience. Future experiences may alter these response tendencies, however, the repertoires that have been acquired are enduring enough to make them appear consistent to a certain degree. Specific situations, however, are associated with specific response tendencies under this view, the strength of which depends on past conditioning (Skinner, 1974). Indeed, Highhouse and Harris (1993) reported that candidates in ACs were rated more

consistently across groups of exercises that assessors considered to be similar to one another. This shows evidence that perhaps similar behavioural repertoires are manifest across AC exercises containing similar environmental contingencies. In these respects, the behavioural paradigm that underlies Lowry's (1997) suggestions for AC construction provides a holistic explanation for, and an acknowledgement of, behaviour in ACs as manifested in situationally specific responses, consistent patterns of behaviour, and/or combinations of these effects. The trait paradigm is neither as lenient nor as comprehensive in its view of AC related behaviour, hence the extensive literature that has attempted to maximise the convergent and discriminant validity of the process.

Much of the reported evidence pervasively suggests that the assessors in ACs are rating candidates in terms of their performance on stand-alone assessment exercises, as opposed to making inferences as to specific ability traits that an individual may possess. ACs continue to be predictive of future performance and particularly promotability. The question therefore remains as to what factors underlie these predictive qualities in the AC process. The data could be interpreted as suggesting that the predictive power of ACs derives from their operation as collections of work samples, that is the exercise effect treated as a source of true variance. There is evidence that work samples hold strong predictive validity in themselves, as previously mentioned.

This approach, proposed herein, is motivated by a multitude of research reporting exercise effects in ACs. This reflects that the traditional trait paradigm does not fit assessor ratings. If behaviourally rated work samples are the underlying driving force behind the predictive power of ACs, then perhaps researchers should be attempting to capitalise on the task-specific notion, rather than making attempts to

infer higher level traits which may have little relevance to, and may even detract from, the predictive qualities of an AC. Such inferences may actually act to impede any efforts to try to improve the AC process. In favour of the trait argument, as Chan (1996) and Klimoski and Brickner (1987) argued, it may be that ACs measure the previously mentioned latent traits (self-efficacy, self-monitoring, and tacit knowledge). These notions may explain the predictive validity of the process.

It is important to note that by choosing to investigate the task-specific AC process, the present study does not seek to reject the trait paradigm. The notion that humans carry relatively stable dispositions that are characteristic and idiosyncratic has a strong empirical basis, and can be found documented elsewhere (e.g., Barrick & Mount, 1991). What the present study does suggest, in relation to an existing body of knowledge (Mischel, 1968), is that situational variables may be powerful and useful determinants of behaviour. When employing managers as assessors, as opposed to psychologists, ACs may well be a measurement device where the inference of relatively stable and enduring traits is inappropriate. Rather, it would be more appropriate to assess individuals in terms of their behavioural responses within a given AC exercise, as the data consistently appear to suggest.

Summary

A robust finding from previous research is the lack of convergent and discriminant validity associated with AC ratings. Although several attempts have been made to improve the construct validity of the AC process, few, if any, of the traditional ACs that employ managers, supervisors or related positions as assessors, have made any notable improvement in explaining assessor ratings in terms of dimensions, as opposed to exercises. Some evidence suggests that employing

psychologists as assessors leads to dimensional ratings and consequently construct validity (Sagie & Magnezy, 1997). However, the habitual use of psychologists as assessors appears unrealistic due to costs. Furthermore, the employment of managers as assessors has been associated with gaining managerial acceptance of the AC process, as well as fostering positive relationships between members of an organisation.

Chan (1996) suggested that if ACs are predictive of stable criteria such as performance and, particularly, promotability, then they must be measuring stable traits that, as yet, remain unidentified. Researchers have suggested that these latent traits may involve such constructs as self-efficacy, tacit knowledge and self-monitoring (Arthur et al., 2000; Chan, 1996; Klimoski & Brickner, 1987). Under a different view, the predictive validity of ACs may reflect that ACs are behaving as a set of stand-alone behaviourally-based work sample exercises (Lowry, 1997). Work samples are supported as being predictive of performance-related criteria due to their relatedness to those criteria (Campion, 1972 Cook, 1998; Muchinsky, 2000). Such an argument applied to ACs would lead to a change in paradigm for the process (i.e., from a dimension-specific model to a task-specific model) where no inference of traits would be made, and the predictor would be highly similar to the criterion behaviour.

Overall Research Aim

The overall research aim in this dissertation is to explore the possibility that AC ratings reflect situationally-specific behavioural responses, rather than trait-based attributions. The first and second studies are to be regarded strictly as minor preliminaries that acknowledge important aspects of the AC measurement issue. The

intention in Study One (see below) is to explore the notion that traits that are not intended for measurement (latent traits) in an AC may account for variance in OARs. Although this concept has been explored in prior research, no prior studies could be found that have investigated the particular set of latent traits that are investigated in Study One. It is reasoned that this study will provide a different perspective to the traditional view of how traits manifest in ACs. The intention in Study Two (see below) is to investigate candidate, and more importantly, assessor perceptions in an AC. This is considered important as an exploratory investigation into the paradigms that assessors adhere to when making AC judgements. Such paradigms may guide the assessor in terms of how they will judge AC candidates.

The final study (Study Three, see below) comprises the main investigation, and directly addresses the overall aim of this dissertation. Studies One and Two, acknowledge that the potential role of trait measurement in ACs should not be abandoned in haste. Study Three, in contrast, focuses on a direct comparison between ACs that treat ratings as situationally specific judgements, and traditional ACs that focus on trait judgements. Although prior studies have investigated the exercise effect and its relationship with criterion measures, no studies could be found that have looked at a direct comparison between a task-specific and dimension-specific AC. The removal altogether of trait categories from ACs and thereby committing to a situationally-specific form of assessment, appears neglected in the research to date.

Hypotheses

Study One, Hypothesis One

Given the consistent findings that ACs are predictive of such stable criteria as performance, but especially criteria related to promotability, Study One looked at the extent to which specific latent constructs may be involved in performance in the AC process. Akin to the suggestions of Chan (1996), Klimoski and Brickner (1987), Lievens (2001b) and Arthur et al. (2000), it was hypothesised that the composite effect (i.e., all constructs entered into a regression) of the constructs: self-efficacy, self-monitoring, and tacit knowledge, would account for a meaningful amount of the variance in OARs in ACs.

Study Two, Additional Research Question

In addition, the present study sought to investigate the extent to which assessors and candidates in ACs in practice perceived that they were being assessed under the dimension-specific or task-specific assessment model. The question surrounded whether assessors thought that they should be assessing stable ability traits, and whether candidates thought that they were being assessed in terms of ability traits, or their performance on stand-alone exercises only. Of greater interest in this question was the view of the assessors, who were expected to hold knowledge of the paradigm under which they were rating as a result of assessor training.

Study Three, Hypothesis Two

Given the robust finding that the exercise effect is robustly salient in non-psychologist assessed AC ratings, it was hypothesised that in a repeated measures

design, dimension-specific and task-specific ACs would share similar psychometric characteristics. Under a traditional trait-based paradigm, the dimension-specific assessment is predicted, in this study, not to display evidence for the measurement of relatively stable and situationally enduring trait-based variables. Rather, exercise effects would prevail, and cast doubt on the notion that stable traits were being measured. However, under a behaviourally-based paradigm, the task-specific ratings would be justified theoretically in terms of displaying behavioural information that is contingent, to some degree, on differing exercises. The isomorphic nature of the psychometric properties associated with the dimension-specific and the task-specific ACs will act as evidence for the notion that the task-based paradigm is theoretically justified, and therefore the more appropriate model of assessment for non-psychologist assessors.

To reiterate, the prediction is made in this study that both the dimension-specific and task-specific approaches will yield similar psychometric characteristics. Both approaches are predicted to show a comparatively large amount of variance associated with participant performance as rated within exercises, i.e., the exercise effect. Variance attributed to participant performance on exercises is conceptually acceptable under the task-specific approach. This is because the behavioural theory that underpins the task-specific approach openly acknowledges and encourages an assessment of situationally specific behaviour. Exercise effects under the trait theory that underpins the dimension-specific approach hold no conceptual justification, and are generally treated as error, evidence against trait measurement, or the result of halo effects. Thus, it is predicted that the evidence from this study will support the contention that a task-specific approach holds greater justification for use by non-psychologist assessors than a dimension-specific approach.

Chapter 2, Study One: Latent Trait Measurement In ACs

Method

Prelude to Studies One and Two

As alluded to in the previous chapter's conveyance of Hypothesis One, this first study sought evidence that the composite effect (i.e., all constructs entered into a regression) of the constructs: self-efficacy, self-monitoring, and tacit knowledge, would account for a meaningful amount of the variance in OARs in ACs. Unfortunately, the sophistication of analyses applied in this first study unavoidably suffers by comparison to that applied later in this dissertation. The simplistic analyses in Study One reflect the weak power provided in the data (e.g., the small size of the usable surviving samples and the manifest imprecision of a key portion of the instrumentation). Nonetheless, some preliminary conclusions are suggested by the notably parsimonious modelling, and their consideration is important prior to unfolding the subsequently more complex final study (Study Three) that follows herein. Moreover, Studies One and Two should be regarded as relatively minor preliminaries, yielding outcomes to guide and justify the more resource intensive approach observed in Study Three.

Military Sample

Participants

Data were collected from a selection board, the military equivalent of an AC, that was already in existence, and was used for recruitment and selection purposes in the Royal New Zealand Air Force (RNZAF) in Auckland, New Zealand. Data were collected at two periods during the year from the 31st of July, 2000 to the 18th of August, 2000, and again from the 2nd of October, 2001 to the 22nd of October, 2001. Assessment ratings were collected from 100 potential recruits. Demographic information for this sample is presented in Table 4.

Assessors

Assessors included 27 male officer-level personnel from the RNZAF located in Auckland, New Zealand. Demographic information for this sample is presented in Table 5. Ethnicity could not be assessed due to confidentiality concerns expressed by the organisation under scrutiny. All assessors had previous experience in assessing participants in multiple selection boards, although none had received any tertiary training in psychology. All assessors had at least two years experience and were regarded as subject matter experts of the position being assessed.

Table 4

Demographic Statistics, Candidates, Study One Military Sample

N = 100	Frequency (= %)
Gender	
Male	83
Female	12
Non Responders	5
Ethnicity	
Caucasian	75
Asian	3
Maori and Pacific Islanders	16
Other	2
Non Responders	4
Age	
15-20	44
21-25	33
26-30	7
31-35	4
Non Responders	12
Education	
No formal education	1
School Certificate	3
Sixth Form Certificate	36
Bursary	32
Bachelor's Degree	14
Higher University Degrees	2
Other	3
Non Responders	9

Table 5

Demographic Statistics, Assessors, Study One Military Sample

N = 27	Frequency	%
Gender		
Male	27	100
Female	0	0
Age		
21-25	6	22
26-30	6	22
31-35	8	30
36-40	2	7
41-45	4	14
51-55	1	4
Education		
Bursary	4	14
Bachelor's Degree	4	14
Masters Degree	10	37
Other	7	26
Non Responders	2	7

The RNZAF Selection Board

The Air Force selection board has been used by the RNZAF since World War II for the purpose of recruitment and selection, and was modelled on the procedures used by the British RAF.

Selection Board Dimensions

Candidates were rated on the following 9 dimensions: Written Communication; Oral Communication; Stability Under Pressure; Relations With Others; Group

Influence/Leadership; Initiative; Determination; Reasoning/Planning and Decision-Making. All dimensions were assessed across all 8 exercises, except for Written Communication, which was only assessed in the written exercise. The following definitions were given, by the RNZAF, for the dimensions assessed in the selection board:

Written Communication: The candidate's ability to express ideas in writing clearly. The ideas are structured in a logical manner with correct spelling and grammar.

Oral Communication: The candidate's ability to orally express ideas with clarity, logical structure, appropriate grammar, pace and non-verbal gestures. This also includes the candidate's ability to effectively listen to others.

Stability Under Pressure: The candidate's ability to perform and achieve tasks under pressure/opposition, or in changing situations. This includes the absence of visible signs of stress.

Relations With Others: The degree to which a candidate is accepted by the group as a team member. The candidate's ability to present him/herself as being pleasant, co-operative, and the ability to get along well with others. The ability to keep the leader informed, reporting problems promptly, and seeking guidance when needed.

Group Influence/Leadership: The candidate's ability to influence others to listen. The ability to enlist support, co-operation and participation. The ability to influence and guide others towards the achievement of the task. The ability to monitor performance, provide positive feedback for effective performance. The ability to assume control in leaderless groups and not be ignored, undermined or reliant on position authority.

Initiative: The candidate's ability to originate ideas and actions to achieve favourable outcomes for the group, seeks opportunity to improve group activity.

Determination: The candidate's ability to apply vigour and drive to tasks. The ability to stay with a position/plan until the objective is reached or unattainable (perseverance); as opposed to being single-minded and continuing in the face of obvious errors. It also includes a sense of urgency.

Reasoning/Planning: The candidate's ability to identify concerns and causes of problems; and find links between information from various sources. The candidate's ability to systematically plan to accomplish tasks; including establishing priorities, time frames and allocating resources.

Decision Making: The candidate's ability to identify a variety of alternatives before selecting a course of action; weighing the advantages and disadvantages; and choosing a logical course of action based on available resources and reasonable assumptions.

Selection Board Exercises

Detailed information relating to the selection board exercises is classified and cannot be reproduced due to reasons of national security. The following brief discussions of the exercise content were authorised by the RNZAF. The following eight exercises were employed to assess the nine dimensions in the selection board:

The Group Discussion: In leaderless groups, candidates were given a topic relating to a current world issue (sourced through the current news media) and were asked to discuss this topic. When the discussion relating to a particular topic became less active, new topics were presented.

Planning Exercise: Involved a simulation exercise where candidates were asked to imagine that they were stranded in a fort in the middle of a desert, and that they had to carry out a planning activity in this situation. Candidates were given five minutes to read about the scenario and 20 minutes to prepare a plan of how they would manage the situation they were in. Again, no leader was assigned to this exercise.

Leaderless Group Exercise: Another leaderless simulation exercise, which involved an outdoor contrived scenario where candidates had to resolve a situation where the aim was to relocate a group to another position situated over a river. Participants were only given two planks of wood to achieve this end.

Chairperson Exercise: In this exercise, each candidate in turn was requested to chair a meeting on a specified topic. Again the choice of topic was at random, although usually reflected a current issue in the media. Five minutes reading time and 10 minutes preparation time were allocated to participants. The chairperson co-ordinated the entire discussion and summarised the issues discussed. Discussions lasted 10 minutes each.

Individual Problem Solving Exercise: A simulation exercise where candidates were set the task of relocating an injured person to a nearby town, and picking up a radio transmitter on the way. This mission was encumbered by a lack of fuel and certain prioritisations and calculations that needed to be completed for the assignment to be successful. Twenty minutes were allocated for reading time, and at the conclusion of the exercise, candidates were asked to show the plan they had developed to the assessors.

Command Situation: A group simulation exercise where each candidate took turns at becoming the leader. Each candidate assigned as the leader was given 10 minutes to complete a task involving the transport of a group over a river using minimal equipment.

Lecturettes: An exercise where candidates were requested to give a two-minute impromptu speech. On a random selection without replacement basis, a candidate was given two topics (e.g., family violence and pollution) and was allocated two minutes preparation and two minutes presentation time for one of these topics.

Group Planning: A leaderless group exercise simulating a peacekeeping effort. Five minutes were allocated for reading time and 20 minutes allocated to discuss. This involved some general problem solving and specific questioning at the conclusion of the exercise.

Written Exercise: Candidates were requested to write an essay on a random topic, so as to assess written communication skills.

Measures

Study One employed three psychological measures to assess the theory that self-efficacy, tacit knowledge and self-monitoring might be involved in successful performance in an AC (Chan, 1996; Klimoski & Brickner, 1987). The following measures were selected for the measurement of these constructs. Please note that the procedure in Study One was repeated for two collection periods, and as such, these measures were used and/or adapted to the requirements of these specific samples.

AC Specific Self-Efficacy: No existing instrument could be located that measured self-efficacy as specifically related to performance on an AC. The measurement of such domain specific self-efficacy was considered vital to the present study, as Bandura suggests that self-efficacy can be construed as a construct that is specific to particular

domains (Bandura, 1997). As such, an extensive enquiry was performed through letter writing to known international researchers in the field of self-assessments related to AC performance (e.g., Halman & Fletcher, 2000). One researcher was identified as having developed a measure of AC domain specific self-efficacy for the assessment of candidate reactions to selection methods (Tovey, 2001). Tovey's AC self-efficacy scale began with a generalised item about the candidate's perceived level of expected performance at the AC. Note that the designation 'extended interview' was given to this particular AC. As such, the first item read: "I believe that I am capable of being successful at this extended interview". This was followed by a series of items that related to performance on the specific exercises in the AC. The first item in this series read: "I believe I will be successful in particular on the following exercises". This was followed by a list of the assessment exercises to create an 8-item measure comprising one general item, and seven exercise related items on a 7-point scale ranging from strongly disagree to strongly agree. Tovey had not collected data with her scale at the time she was contacted, and as such, no psychometric information pertaining to the scale was available. Tovey's scale was, however, intuitively appealing as a face valid scale framework that could easily be adapted to different ACs. As a result, Tovey's framework was employed in this sample.

General Self-Efficacy: Several researchers have hypothesised that a global sense of self-efficacy could result from several self-efficacy fostering or diminishing experiences across different domains. Labelled general self-efficacy, this construct asserts that a collection of experiences related to varying levels of self-efficacy in the past could carry into perceived self-efficacy expectations in new situations for an individual. Most of the

current research into general self-efficacy has focused on a scale developed by Sherer, Maddux, Mercandante, Prentice-Dunn, Jacobs, and Rogers (1982) and later researched and revised by Woodruff and Cashman (1993) and Bosscher and Smit (1998). The present study utilised the version of the Sherer et al. scale that was presented in Bosscher and Smit (1998), which comprised a 12-item general self-efficacy scale (GSES-12). The scale breaks down general self-efficacy into 3 sub-constructs (initiative, effort and persistence) and also purports to measure a higher order general self-efficacy construct composed of the combination of these three components. Various studies, in general, have found acceptable levels of internal consistency for the general self-efficacy scale. Minor changes were made to some of the items in the scale across the different studies. For the overall general self-efficacy scale, Cronbach alpha reliability coefficients of .86 and .69 were reported by Sherer et al. (1982) and Bosscher and Smit (1998) respectively. For the subscales of the general self-efficacy scale, Woodruff and Cashman (1993) and Bosscher and Smit (1998) found the following Cronbach alpha coefficients for the three scales respectively: Initiative: .74; .64, Effort: .75; .63, Persistence: .64; .64. These inter-item consistency coefficients fall within the limits of moderate acceptability as suggested by Nunnally and Bernstein (1994). Internal consistency for the Bosscher and Smit (1998) study was slightly lower than the other studies. This may have been due to the fact that Bosscher and Smit excluded 5 items that were found in a pilot study to have low item-total correlations and ambiguous wording. The alpha differences might also have been due to Bosscher and Smit's use of elderly people as a sample, while the studies by Sherer et al. and Woodruff and Cashman employed student participants. In any case, it was decided that the internal consistency estimates for the Bosscher and Smit version of the

scale were still within the limits for acceptability, and that a slightly lower number of items might assist to maximise return rates. Note that only the unitary scale was employed in the present study. In this study, General self-efficacy was measured on a 7-point scale ranging from 1 (disagree strongly) to 7 (agree strongly).

Convergent validity evidence has been reported by Sherer et al. with the finding that general-self-efficacy, as measured by the general self-efficacy scale, correlated positively with the likelihood that a given individual was in current employment, with quitting from fewer jobs and being fired from fewer jobs, with educational level and military rank. General self-efficacy was also found to correlate with an internal locus of control and self esteem. Woodruff and Cashman (1993) found a similar pattern of correlational data, with positive relationships found between general self-efficacy and personal mastery, task specific self-efficacy, and expectations of receiving higher grades.

Tacit Knowledge: Klimoski and Brickner (1987) specifically theorised that a type of managerial intelligence might be related to the extent to which an individual could be successful in an AC, and on later criterion measures of performance and promotability. As such, the present study employed the Tacit Knowledge Inventory for Managers (TKIM) (Wagner, 1985). Although the participants were not managers themselves, the current measurement could be viewed as an indication of an individual's aptitude for being a successful manager. The measure could also be viewed as a gauge of the extent to which individuals already held the characteristics that may be conducive to holding managerial intelligence that theoretically may in turn assist them towards success in the AC and into the job. Using the TKIM on non-managerial samples is certainly not

unprecedented, and it has been used successfully for the assessment of non-managerial individuals in past research (Wagner & Sternberg, 1985).

Colonia-Willner (1998) reported Cronbach alpha coefficients of .85, .83 and .85 and Wagner and Sternberg (1991), the coefficients .74 and .80 for separate samples for the entire TKIM scale. The theory relating to managerial tacit knowledge delineates the construct into various components relating to managerial intelligence concerning self; others; and tasks. Colonia-Willner found moderate internal consistency coefficients for these sub-constructs with respective Cronbach alphas of .74, .67 and .64 in her first study, .70, .64 and .60 in her second study and .74, .68 and .65 in her third study. This might suggest that the TKIM may be better employed as a unitary scale. In the interests of maximising measurement precision, it was decided, on the basis of the previously mentioned study, to employ the unitary conceptualisation of this construct in Study One.

A multitude of evidential information exists for the convergent and discriminant validity of the TKIM, some of which has already been discussed in the previous section, and as such, this will be only briefly mentioned here. Discriminant validity studies suggest that tacit knowledge is independent of academic performance and cognitive ability test scores (Colonia-Willner, 1998; Wagner & Sternberg, 1991), and convergent evidence suggested that scores on the TKIM were related to job performance (Wagner & Sternberg, 1985). Colonia-Willner found that the best scorers on the TKIM were more experienced managers, which corresponds to the theory that tacit knowledge is gleaned from experience.

Two versions of the TKIM exist. One of these employs expert samples to create deviation scores for scoring participants. The present study employed a version of the

TKIM that does not require the use of an expert sample (Wagner, 1985), in the interests of time and available resources. The number of items in the total scale for this version was 39. These 39 'real' items were imbedded within another 127 dummy items that were not scored. A set of items related to a set of 12 managerial scenarios that were each presented in a vignette. These items came as a booklet sent directly from the author (Wagner, 1985) and were presented on a 7-point scale ranging from 1 (not important) to 7 (extremely important). The scale broke managerial tacit knowledge into the areas of tacit knowledge related to managing one's career, managing self, managing others, and "other" items that were described as discriminating between those who had higher levels of tacit knowledge, but did not fit the theory. Wagner replaced the 'career' scale with tacit knowledge relating to 'tasks' in a later version of the tacit knowledge inventory for managers (a version that requires the use of expert samples) in response to a subtle development in the theory of the tacit knowledge concept (Wagner and Sternberg, 1991). In any case, just as the evidence suggests for the version of the TKIM that employs expert samples, the results of Wagner and Sternberg's (1985) article suggested that the non-expert sample version of the scale should be viewed as a measure of a unitary tacit knowledge construct, with one study showing evidence of moderate levels of acceptable internal consistency for the entire measure (at .68).

Self-monitoring: The present study employed the 12-item O'Cass (2000) revision of the Lennox and Wolfe (1984) Revised Self-Monitoring Scale. The O'Cass revision was a subtle modification of the scale, whereby one item was dropped from the original measure because a pilot study revealed poor reliability and item total correlations, and the

scale poles were changed from a 6-point scale ranging from 1 (certainly always false) to 6 (certainly always true) to a new 6-point scale ranging from 1 (strongly disagree) to 6 (strongly agree). It was decided to use the latter of these poles, as O'Cass found that participants were better able to interpret the modified scale.

Lennox and Wolfe (1984) conceptualised self-monitoring as being composed of two underlying factors: self-monitoring ability and self-monitoring sensitivity. The revised self-monitoring scale reflects this theory by attempting to tap both of these factors. O'Cass found Cronbach alpha coefficients of .86 and .85 for the two subscales measuring self-monitoring ability and self-monitoring sensitivity respectively. For the entire scale, the reported Cronbach alpha was .87. This study also found convergent relationships between high scores on the self-monitoring scale and concern for personal image.

OARs: OARs were derived from the average of two separate OARs specified by two independent senior assessors. To elaborate, according to Air Force policy, upon completion of the assessment exercises, two senior officers decided upon two independent OARs based on their judgement, the assessment ratings, and their observations during the entire assessment process. The OARs themselves were on a four point scale with the anchors A (strongly recommended), B (recommended), C (marginal), and D (not recommended). As these categories were intended, according to airforce officals, to graduate from high to low, they were treated numerically as A (4), B (3), C (2), and D (1) for the purposes of analysis.

Procedure

As the theory suggested that the three constructs self-efficacy, tacit knowledge and self-monitoring in their combination may contribute to the effective performance in ACs, the design was set up so that measurements of the constructs were taken before the AC. Potential participants were invited to partake in the present research prior to their arrival at the AC, by sending questionnaires along with information packs that the RNZAF administered through the post. All questionnaires were coded, and were sent via the post directly back to the researcher. Ratings of individual's performance on the AC were also collected once the AC had been completed. The codes for pre-measure constructs and AC measures were then matched for subsequent analysis. Note, this procedure was repeated in the same manner for the sample described below.

Organisational Sample

The organisational sample for Study One was a repeat of the study described above for the military sample. As such, the measures and procedure were identical across both samples. The participants, assessors, and key aspects of the AC in the organisational sample are described below.

Participants

For Study One, data were collected from an AC that was already in existence and was being used for recruitment and selection purposes by a large retail company in Bayfair, Tauranga, New Zealand. Data were collected from the AC, which ran for one week, beginning on the 14th of August and ending on the 21st of August, 2001. AC

ratings were collected from 87 potential recruits. Demographic information on this sample is presented in Table 6.

Table 6

Demographic Statistics, Candidates, Study One Organisational Sample

N = 87	Frequency	%
Gender		
Male	21	24
Female	66	76
Ethnicity		
Caucasian	54	62
Asian	6	7
Maori	18	21
Other	3	3
Non Responders	6	7
Age		
15-20	8	9
21-25	14	16
26-30	5	6
31-35	9	10
36-40	10	11
41-45	7	8
46-50	12	14
51-55	13	15
56-60	7	8
66-70	1	1
Non Responders	1	1
Education		
No formal education	13	15
School Certificate	23	26
Sixth Form Certificate	16	18
Bursary	2	2
Bachelor's Degree	5	6
Higher University Degrees	3	3
Other	25	30

Assessors

Assessors included 17 managerial level staff members of the retail organisation from various parts of New Zealand. Only partial demographic information was available from the assessor group due to non-response. Of the seven who responded, three were male, four were female and their mean age was 21.86 ($SD = 2.80$). All were located in Auckland, New Zealand. According to information subsequently obtained from the organisation, the non-responding assessors were older and were more experienced than those who did respond to the demographic items. Also according to information obtained from the organisation, all assessing participants had previous experience in assessing ACs for the retail store under scrutiny. Only one of the assessors had previously received any training in psychology, having completed a Bachelors degree. All participants had over two years experience in their positions, and were regarded as subject matter experts of the position being assessed.

The AC

An external multi-national consulting company constructed the AC under scrutiny for the purposes of recruitment and selection. Rather than a bespoke approach, the consulting company who designed the AC selected 'off-the-shelf' competencies that were deemed relevant, and assessed these through 'off-the-shelf' exercises that were also deemed relevant for assessment.

AC Dimensions

Candidates were rated on the following 9 dimensions: Interpersonal Skills; Social Confidence/Assertiveness; Problem Solving/Decision Making; Decisiveness; Results Focused/Perseverance; Customer Focus; Team Player; Sales; Mentoring. One other dimension called Numeracy was also assessed through an external paper and pencil test and a single dichotomous pass or fail rating. Two other dimensions, named Availability and Personal Presentation, were again dichotomous items, which probed whether the individual was available to perform the position, and whether their personal presentation was up to standard, respectively. These last three dimensions were not included in the present study as they did not form part of the psychological assessment process of the AC, and these factors did not utilise multitrait-multimethod assessment methodology (see Figure 1). Different dimensions were assessed across exercises as outlined in the exercise competency matrix in Figure 1. Note that blackened areas in Figure 1 indicate where a dimension was not assessed. The following definitions were provided for the other dimensions in the AC:

Problem Solving/Decision Making: solves difficult problems with effective solutions; asks good questions and probes for answers; looks beyond the obvious; able to consider information from a variety of sources; exercises good judgement when making decisions; comes up with new and innovative ideas; sees the long-term impact of decisions; has good sound judgement about which creative ideas and suggestions will work; brings creative ideas of others to the fore.

COMPETENCY/ EXERCISE	Numeric ability	Egg exercise	Group Interview	Lost at Sea	Availability	Final Rating
Problem-Solving						
Decisiveness						
Goal Orientation						
Interpersonal Skills						
Social Confidence						
Numeracy	Y N					
Availability					Y N	
Team Player						
Customer Service						
Sales						
Mentoring						
Personal Presentation		Y N				

Figure 1. Competency/Exercise Matrix for Study One, Organisational Sample.

Decisiveness: makes timely business decisions based on assessment of facts, assumptions and implications; makes timely decisions, sometimes with incomplete information and under tight time pressure; most solutions turn out to be correct and accurate when judged over time; has a bias for action.

Goal Orientation: can be counted on to reach goals successfully; very bottom line orientated; pushes self and others to achieve results; pursues goals with energy and drive; seldom gives up without finishing, especially in the face of setbacks; is resourceful and tenacious in finding an alternative means to reach a goal.

Interpersonal Skills: communicates well with all kinds of people internally and externally; builds appropriate rapport; builds constructive and effective relationships; uses diplomacy and tact; practices active listening; has the patience to hear people out; is easy to approach and talk to; puts others at ease; genuinely cares about others; is available

and ready to help; acknowledges others' concerns; is co-operative; gains the trust and respect of peers; works with others, sharing tasks and accountabilities.

Social Confidence/Assertiveness: seeks out social situations and interacts confidently in group situations; can challenge others' views appropriately; comfortable sharing own perspective with managers and peers in a group situation; stands up for what he or she believes in and holds own ground, even in the face of opposition.

Team Player: invites input from each person and shares ownership and visibility; makes each individual feel as though their work is important; is someone people like working with; creates strong morale and spirit in the team; shares successes; fosters open dialogue; creates a feeling of belonging in the team; works co-operatively with others.

Customer Service: is dedicated to meeting the expectations and requirements of internal and external customers; gets first hand customer information and uses it for improvements in products and services; talks and acts with customers in mind; establishes and maintains effective relationships with customers and gains their respect and trust.

Sales: understands and can describe the steps in the sales process; understands the importance of sales; acts with the customer in mind at al times.

Mentoring: first identified how much the subject knew; created a plan (not necessarily written) to use to develop the person; used an appropriate approach(es); identified follow-up action.

OARs: AC overall ratings constituted the average ratings across all of the dimensions assessed. The mechanical integration of ratings was used in congruence with the practice of the organisation whose AC was under study. Such methods of integration have been

deemed acceptable according to the latest international guidelines for ACs (International Task Force on Assessment Center Guidelines, 2000). Each dimension was rated on the following scale, ranging from 1 (The person does not have the competency), 2 (Individual does not have the competency level required), 3 (Individual does not quite have the competency required), 4 (Individual has the required competency level), to 5 (Level of competency is beyond that which the position requires).

AC Exercises

The following three exercises, designed by an external consulting company, were employed to assess the nine dimensions in the AC, along with their descriptions:

Egg Simulation Exercise: This exercise comprised a low-fidelity teamwork activity, where participants set about constructing a framework composed of certain stationery items (e.g., paper, paper-clips, a balloon, string). The object of the activity was for the group to construct a framework that would allow an egg to be dropped from a height of approximately two metres onto a hard surface, such that the egg did not break.

Group Interview: This comprised an low-fidelity individual exercise, contextualised within a group setting. Each individual in a group was asked a series of questions to which they had to formulate an answer. Example questions included 'What is the most challenging thing you have done, and what did you learn from it?' and 'What is the most rewarding experience you have had in a team? Why was this rewarding and what made it different from other team experiences?'

Lost At Sea Simulation Exercise: This comprised a low-fidelity teamwork exercise, where participants were requested to imagine that they were adrift in a private yacht, irreparably damaged by a fire of unknown origin. The group were told that certain items had remained intact, and that, as a group, they were to rank these items in terms of their overall importance to survival.

Note that these were all low-fidelity simulation exercises, however this is typical of many of the AC simulations offered by consulting companies worldwide (Muchinsky, 2000). This is also in agreement with the international guidelines for AC development, as these guidelines stipulate that the fidelity of simulation exercises may be relatively low if the centre is used for early identification and selection programs and for non-managerial personnel (International Task Force on Assessment Center Guidelines, 2000). The AC employed in Study Three also fitted both of these criteria.

Results

The data from two separate samples, one from the Royal New Zealand Air Force (RNZAF) section of the New Zealand military, the second taken from a large New Zealand based departmental retain chain, were explored as outlined below. The following statistical considerations were applied in the analysis. Power analyses were conducted, followed by the calculation of relevant descriptive statistics, comprising means and standard deviations for each measure. Bivariate correlations between variables and internal consistencies for each measure were calculated. Multiple

regression analyses were conducted to investigate the extent to which the composite of the variables under study (self-efficacy, self- monitoring and tacit knowledge) explained meaningful variance in OARs. In the military sample, the amount of variance associated with the full composite theoretically explained approximately 16% of the variance in OARs in the population. When correcting for validity shrinkage for generalisation across samples, the full composite explained approximately 1% of the variance in OARs. The strongest predictor in this composite was tacit knowledge, despite the fact that the tacit knowledge measure held low internal consistency in this sample. In the organisational sample, the results suggested that the composite measures explained very little variance in OARs (approximately 4% of population variance).

Military Sample

Practical problems occurred with the measurement of tacit knowledge in the military sample due to non-response on the tacit knowledge inventory. The sample was divided into two separate runs of the AC over two time periods. Initially, the participants were administered a version of the inventory that required the use of an expert sample from which deviation scores would be calculated. Unfortunately, the expert sample had such a high non-response rate, that the questionnaire had to be abandoned. On the second run of the AC, the sample was administered a version of the questionnaire that did not require the use of an expert sample. Thus, the analyses will be divided into two sets. Set One will show the entire sample with the measure of tacit knowledge removed ($N_1 = 100$). Set Two will show the second run of the AC only, where the tacit knowledge inventory is included ($N_2 = 44$).

Set One

Set One utilised the full military sample ($N = 100$). The total pool of people who applied for the group of positions was 116, thus 100 was a high response rate at roughly 86%. Data were imputed for missing values using EM (expectation maximisation), which employs an iterative process by which to estimate missing values. This method was recommended by Gold and Bentler (2000) for optimal data substitution, regardless of sample size, proportion of missing data and distributional characteristics. Imputations were required for 16% of OARs, 14% of specific self-efficacy ratings, 0% of general self-efficacy ratings and 26% of self-monitoring ratings. In general, the response rates for the questionnaires were reasonably high, except perhaps for the self-monitoring ratings.

After imputing the OARs with EM (as stated above), the two sets of ratings provided by the senior assessors were found to correlate at $r = .93, p < .01$. This suggests that the assessors were generally in agreement with one another on their derivation of OARs. Note that all power analyses in this study were conducted using GPOWER version 2.0 (Faul & Erdfelder, 1992). An a priori power analysis was performed for multiple regression analyses for three predictors. This analysis revealed that the number of cases in this study was more than the 77 cases necessary for a 2-tailed test at the .05 level of significance, at a power level of .80 for medium effect sizes. The current analysis therefore achieved acceptable power, contingent on obtaining medium level effect sizes. Note that GPOWER converts the Cohen (1988) measure of effect size (f^2) which is, by convention, set at 0.15 for medium effect size, into an estimate of multiple R^2 medium effect size (see Murphy & Myors, 1998 for a summary on effect size conventions). This principle applies to all studies within this chapter.

Table 7 shows the means and standard deviations for the measures used in the study. Particularly with respect to the specific self-efficacy scale, (SSE) small standard deviations could represent range restriction problems, as a small amount of systematic variance in scores can make it difficult for correlations to manifest. Table 8 shows the bi-variate correlations and internal consistencies for the measures used in the study. All of the internal consistency coefficients were within the limits suggested by Nunnally and Bernstein (1994, p. 252). Significant correlations were found between GSE and SSE ($r = .48, p < .01$). This was the strongest relationship with respect to magnitude. Similarly, it was again found that SSE was related to SM ($r = .23, p < .05$). The pattern of correlations and lack of significance between the OAR and the set of presumed predictors suggests no bi-variate relationship. As the strongest correlations were among the set of presumed predictors, the possibility of

Table 7

Overall Means and Standard Deviations for Measures Employed in Set One of the Military Sample

Scale	*M*	*SD*
Overall Assessment Ratings (OAR)	2.32	0.94
AC Specific Self-Efficacy (SSE)	5.84	0.10
General Self-Efficacy (GSE)	6.10	1.11
Self-Monitoring (SM)	4.46	1.00

Table 8

Bivariate Correlations Between Measures Employed in Set One of the Military Sample

Scale	1	2	3	4
1. OAR	(.96)[a]			
2. SSE	.10	(.79)		
3. GSE	-.07	.48**	(.76)	
4. SM	.17	.23*	.13	(.73)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to two items that made up the overall rating. The reader is cautioned that only the correlation between GSE and SSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

confounding exists, and therefore multivariate analysis was employed. The bi-variate correlations may indicate a violation of multicollinearity assumptions, particularly with respect to the relationship between GSE and SSE. The reader is cautioned that only the correlation between GSE and SSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

For the reasons detailed earlier, taken together with the restrictive sample size of the present study, standard all-in regression was selected as the multivariate technique that would be most appropriate. The summary statistics in Table 9 display two indices of R^2 adjusted for validity shrinkage (Rosenthal & Rosnow, 1991). For a detailed account of these indices and their respective formulae, the reader is directed to Bobko (1990). The first index, labelled 'adjusted $R^{2'}$, estimates what would happen if the sample, in a given study, were to be the population in its entirety. The second

index is labelled 'shrunken R^{2}', and estimates how well a given model would predict in other future samples on average (i.e., in the population of samples) (Bobko, 1990).

The adjusted R^2 suggested that if the sample were the population, the set of predictors theoretically accounted for 2.1% (*ns*) of the variance in the OARs (see Table 9) (Licht, 1995). The shrunken R^2 suggested that in other samples, the set of predictors would account for .02% (*ns*) of the variance in OARs. None of the predictors displayed significant partial relationships with the criterion. Note that a post-hoc power analysis revealed for an $R^2 = .051$, a sample size of 207 would be needed to achieve power of .80 with three predictors in a regression model. In the case of this study, power was equal to 0.45, and thus the probability of making a type II error was .55 (Rosenthal & Rosnow, 1991). The present study, therefore, stacked

Table 9

Multiple Regression Analysis for the Prediction of OARs in Set One of the Military Sample

		Partial Regression Weights	
Predictors	Raw (B)	Standardised Beta	95% Confidence intervals
SSE	0.02	.14	-.02 < B < .07
GSE	-0.02	-.15	-.05 < B < .01
SM	0.03	.16	-.01 < B < .06
Intercept	1.17		

Summary: $R = .225(ns)$, $R^2 = .051$, Adjusted $R^2 = .0213$, Shrunken $R^2 = .0002$

the odds in favour of the null hypotheses, given the attenuated effect size and small sample size.

A residual analysis revealed no clear threat to homoscedesticity assumptions. Evidence assuaging multicollinearity concerns was found with variance inflation factor (VIF) indices being less than 10 (maximum = 1.35, minimum = 1.05) (Chatterjee, Hadi & Price, 2000), and tolerance indices did not approach zero (minimum = 0.74, maximum = 0.95) (Tabachnick & Fidell, 1983). Additionally, eigenvalues did not differ greatly (maximum = near zero, minimum = near zero) (Belsley, et al., 1980). The scores were not normally distributed, with a large cluster to the negative side and two clusters toward the centre of the distribution. This may reflect an overuse of central gradings. The residual plots also showed evidence of several outliers. Reconsideration of these distributional problems did not alter conclusions regarding the non-significant outcomes. Nonparametric significance tests yielded similar outcomes regarding the view of linear relations between all measures (see Table 10), and inspections of scatterplots revealed no reason to suspect curvilinear outcomes.

Note that the bivariate correlations between the presumed predictors and OARs, corrected for attenuation due to unreliability[b] (Schmidt & Hunter, 1996, p. 201) were as follows. SSE and OAR (r = .11, *ns*); GSE and OAR (r = -.08, *ns*); SM and OAR (r = .20, *ns*). The corrected correlations here were considered to be similar to those correlations reported in the uncorrected bivariate correlations, thus no further correctional analyses were conducted. Note that the corrected coefficients may have

[b]Bobko (2001) asserts that it is "customary to test the original, uncorrected Pearson r for statistical significance and then report corrected r as the best point estimate of the true relationship between the variables" (p. 82). This is because the t tests associated with Pearson's r assume that the sample-based r is computed. This principle is applied to all corrected correlations within this chapter.

Table 10

Spearman's Rho Between Measures Employed in Set One of the Military Sample

Scale	1	2	3	4
1. OAR	(.96)[a]			
2. SSE	.08	(.79)		
3. GSE	-.07	.48**	(.76)	
4. SM	.13	.22*	.17	(.73)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to two items that made up the overall rating. The reader is cautioned that only the correlation between GSE and SSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

been higher than reported here, had a proper index of interrater reliability been available for OARs. Schmidt and Hunter (1996, p. 209) describe Cronbach's alpha as an estimate of *intrarater* reliability in contexts such as these. Such measures tend to give higher estimates when compared to what might be expected from other indices of interrater reliability. Unfortunately, traditional intraclass correlation-based indices of interrater reliability (e.g., Shrout & Fleiss, 1979) could not be employed in this sample because specific information relating to the allocation of assessors was not provided by the organisation under study. The other samples in Study One were also afflicted with this potential limitation.

Set Two

Set Two included the same data as above, but only selected those cases that integrated the TKIM (n = 44). Bearing in mind that the total number of people who applied for the group of positions in this sample was 46, a sample of 44 was thought to constitute a high response rate at roughly 96%. OARs considerations were the same in Set One as for Set Two. An a priori power analysis was performed for multiple regression analyses with four predictors. This analysis revealed that the number of cases was less than the 85 cases necessary for a 2-tailed test at the .05 level of significance, at a power level of .80 for medium effect sizes. The current analysis therefore stacked the odds in favour of the null hypothesis, contingent on obtaining medium level effect sizes.

Table 11 shows the means and standard deviations for the measures used in the study. Restricted range may have been a problem, particularly for the OAR in this sample, which yielded a relatively small standard deviation. This may have restricted

Table 11

Overall Means and Standard Deviations for Measures Employed in Set Two of the Military sample

Scale	*M*	*SD*
OAR	2.38	0.10
SSE	5.77	1.16
GSE	5.63	1.45
SM	4.48	0.94
Tacit Knowledge Inventory for Managers (TKIM)	3.95	1.73

opportunities for correlations to manifest. Table 12 shows the bi-variate correlations and internal consistencies for the measures. The internal consistency coefficients for SM were bordering on the lower end of those suggested by Nunnally and Bernstein (1994, p. 252) and the coefficient for the TKIM was well below the suggested limits. The measurement of tacit knowledge in this portion of the study therefore lacked internal consistency. The difference in terms of internal consistency and bi-variate correlation between the TKIM as measured in the organisational sample (discussed later) and the military sample was probably due to sampling error (at $n = 44$), however, it may be due to real differences between the samples. This is discussed further in the discussion section.

Table 12

Bivariate Correlations Between Measures Employed in Set Two of the Military Sample

Scale	1	2	3	4	5
1. OAR	(.95)[a]				
2. SSE	.12	(.80)			
3. GSE	-.04	.47**	(.73)		
4. SM	.36*	.24	-.07	(.62)	
5. TKIM	.36*	-.01	.04	.09	(.41)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to two items that made up the overall rating. The reader is cautioned that only the correlation between GSE and SSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

The predictive validity coefficient relating to the TKIM, that is, the correlation between the presumed predictor TKIM and the DV OAR, is possible despite the low reported internal consistency. According to Bobko (2001), a rule of thumb concerning correlations and their relationship to internal consistency is that the predictive validity of a measure can be no greater than the square root of its reliability. The reported correlation between TKIM and OAR, at .36, is less than the square root of the internal consistency estimate of the TKIM's reliability (.64).

All conclusions from here on must be drawn bearing in mind the implications of low reliability in the measure of tacit knowledge. These include that it is questionable that the measure was actually measuring a unitary concept, and it may be that its respective components were sufficiently unrelated to call their union into question, for this particular sample. Thus, the respective components of the TKIM did not appear to share sufficient dimensionality in this sample. Given this caution, Table 12 showed significant correlations between GSE and SSE ($r = .47, p < .05$), as across all runs of this study. Of greater interest was the finding that SM ($r = .36, p < .05$) and TKIM ($r = .36, p < .05$) both displayed significant correlations with the DV, OAR. The reader is cautioned further that only the correlation between GSE and SSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

The strongest bi-variate relationship here was between two of the presumed predictors. On this occasion however, two of the presumed predictors displayed relationships with the DV. This would suggest, again, the need for multivariate analysis. For this reason, and with respect to the restrictive sample size, standard all-in regression was selected as the appropriate technique. The adjusted R^2 suggested that the set of predictors in Table 13 theoretically accounted for 16% ($p < .05$) of the

Table 13

Multiple Regression Analysis for the Prediction of OARs in Set Two of the Military Sample

		Partial Regression Weights	
Predictors	Raw (B)	Standardised Beta	95% Confidence intervals
SSE	0.01	.01	-.04 < B < .07
GSE	-0.01	-.07	-.06 < B < .04
SM	0.06	.30	.00 < B < .13
TKIM	-0.03	.33*	.01 < B < .06
Intercept	-5.48		

Summary: R = .489*, R^2 = .239, Adjusted R^2 = .1609, Shrunken R^2 = .0108

* $p < .05$ (2-tailed)

variance in OARs in the population. This result was, with respect to overall magnitude, seemingly stronger than the effects found in the previous example. The TKIM was the strongest predictor in this regard, which was notably intriguing, given its lack of internal consistency. The shrunken R^2 suggested that across different samples, the set of predictors would account for 1.08% ($p < .05$) of the variance in OARs. Note that a post-hoc power analysis revealed for an effect size of R^2 = 0.239, a sample size of 44 would be needed to achieve power of .80 with four predictors in a regression model. In the case of this study, power was equal to 0.81, and thus the probability of making a type II error was 0.19 (Rosenthal & Rosnow, 1991).

A residual analysis revealed no salient threat to homoscedesticity assumptions. Evidence assuaging multicollinearity concerns was found with VIF indices being less than 10 (maximum = 1.43, minimum = 1.02) (Chatterjee, et al., 2000), and tolerance indices did not approach zero (minimum = 0.70, maximum = 0.99) (Tabachnick & Fidell, 1983). Additionally, eigenvalues did not differ greatly (maximum = near zero, minimum = near zero) (Belsley, et al., 1980). The distribution of scores did not fit a perfect normal curve with a large cluster of scores toward the positive end of the distribution. This may again reflect an overuse of mid to upper gradings. The residual plots also showed evidence of minimal outliers. Reconsideration of these distributional problems did not alter conclusions regarding the non-significant outcomes. Nonparametric significance tests yielded similar outcomes regarding the view of linear relations between all measures (see Table 14), and inspections of scatterplots revealed no reason to suspect curvilinear outcomes.

Table 14

Spearman's Rho Between Measures Employed in Set Two of the Military Sample

Scale	1	2	3	4	5
1. OAR	(.95) [a]				
2. SSE	.13	(.80)			
3. GSE	-.03	.51**	(.73)		
4. SM	.25	.29	.03	(.62)	
5. TKIM	.34*	-.08	.03	-.01	(.41)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to two items that made up the overall rating. The reader is cautioned that only the correlation between GSE and SSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

The bivariate correlations between the presumed predictors and OARs, corrected for attenuation due to unreliability (Schmidt & Hunter, 1996, p. 201) were as follows. SSE and OAR (r = .14, *ns*); GSE and OAR (r = -.05, *ns*); SM and OAR (r = .47, $p < .05$); TKIM and OAR (r = .66, $p < .05$). The correlations between OARs and SM and between OARs and TKIM were stronger than correlations reported in the uncorrected bivariate correlations. This was because the corresponding measures, particularly the TKIM, were afflicted with low internal consistency. As previously mentioned, the corrected coefficients may have been higher than reported here, had a proper index of interrater reliability been available for OARs (Schmidt and Hunter, 1996, p. 209).

Supplementary Analysis for Set Two of the Military Sample

Given the enlarged bivariate correlations observed when correcting for unreliability in Set Two, it was decided that as a supplementary analysis, problematic items would be removed from the TKIM in order to improve its internal consistency, whilst preserving its construct domain coverage. Murphy and Davidshofer (2001) suggest that test items should be representative of the domain of attributes being measured. As mentioned in the method section, the TKIM covers tacit knowledge relating to managing career, self, other people, and 'other' discriminating items. Item analyses revealed several negative item-total correlations in the data for Set Two. Negative item-total correlations indicate divergence between particular items and test scores or, of course, the possibility of encoding errors or the possible need for reverse coding, et cetera (Murphy & Davidshofer, 2001). Items with negative or low item-total correlations were removed selectively to maintain the theoretical framework of the TKIM to the greatest degree possible. In the original scale, fifteen items related

to managing career in the TKIM, sixteen related to managing self, four items related to managing other people and four items were classified as 'other'. In order to assist in maintaining construct domain coverage, items were not removed from the managing other people and the 'other' scales. From the managing career scale, four items with negative item-total correlations were removed (most divergent item-total correlation = -.42, least divergent item-total correlation = -.11). From the managing self scale, nine items with negative or low item-total correlations were removed (most divergent item-total correlation = -.21, least divergent item-total correlation = .08). Of the thirteen items removed in this military sample, nine of the TKIM items also loaded negatively in the organisational sample (described later) in Study One. The analyses for Set Two of the military sample in Study One were repeated with the altered version of the TKIM.

The grand mean for the revised TKIM was 3.82 (SD = 1.71). This lack of variation may have lead to problems related to range restriction, and thus, may have restricted the extent to which correlations manifested in this sample. Table 15 shows the bi-variate correlations and internal consistencies for the measures employed. Note that the correlations are identical to those displayed in Table 12, except that, most notably, the relationship between TKIM and OAR increased from .36 ($p < .05$) to .42 ($p < .01$) in this supplementary analysis. This correlation was between a presumed predictor and the DV, OAR. Correlations were observed between the set of presumed predictors, and thus, the possibility of confounding existed. The reader is cautioned, as with the initial analysis of Set One, that the bi-variate correlations may indicate a violation of multicollinearity assumptions, particularly with respect to the relationship between GSE and SSE. The reader is cautioned that the correlations between GSE and SSE, and between TKIM and OAR manifest as statistically significant outcomes

Table 15

Bivariate Correlations Between Measures Employed in Supplementary Set Two of the Military Sample

Scale	1	2	3	4	5
1. OAR	(.95)[a]				
2. SSE	.12	(.80)			
3. GSE	-.04	.47**	(.73)		
4. SM	.36*	.24	-.07	(.62)	
5. TKIM	.42**	.08	.17	.02	(.73)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to two items that made up the overall rating. The reader is cautioned that only the correlations between GSE and SSE and between TKIM and OAR manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$). Note that inter-item consistency was identical to those displayed in Table 12, except that the revised TKIM scale yielded a Cronbach's alpha of .73, which was within the limits suggested by Nunnally and Bernstein (1994, p. 252).

For the same reasons outlined in Set One, Study One, standard all-in regression was selected as the multivariate technique that would be most appropriate. The summary statistics in Table 16 display two indices of R^2 adjusted for validity shrinkage (see Set One, Study One for a brief description). The adjusted R^2 suggested that the set of predictors in Table 16 theoretically accounted for 23% ($p < .01$) of the

Table 16

Multiple Regression Analysis for the Prediction of OARs in Supplementary Set Two of the Military Sample

		Partial Regression Weights	
Predictors	Raw (B)	Standardised Beta	95% Confidence intervals
SSE	0.01	.06	-.04 < B < .06
GSE	-0.02	-.12	-.07 < B < .03
SM	0.07	.32*	.01 < B < .13
TKIM	0.03	.42**	.01 < B < .05
Intercept	-8.52		

Summary: R = .550**, R^2 = .302, Adjusted R^2 = .2304, Shrunken R^2 = .0284

* p < .05 (2-tailed) ** p < .01 (2-tailed)

variance in OARs in the population. This result was, with respect to overall magnitude, seemingly stronger than the effects found in the Set One. Thus, it is likely that the lack of reliability in the TKIM attenuated potential relationships with OARs. The TKIM was the strongest predictor in this model, coupled with SM, which also reached significance as a single predictor. The shrunken R^2 suggested that in different samples, the set of predictors would account for 2.84% (p < .01) of the variance in OARs. Note that a post-hoc power analysis revealed for an effect size of R^2 = 0.302, a sample size of 33 would be needed to achieve power of .80 with four predictors in a regression model. In the case of this study, power was equal to 0.93, and thus the probability of making a type II error was 0.07 (Rosenthal & Rosnow, 1991).

A residual analysis revealed no obvious threat to homoscedesticity assumptions. Evidence assuaging multicollinearity concerns was found with VIF indices being less than 10 (maximum = 1.42, minimum = 1.03) (Chatterjee, et al.,2000), and tolerance indices did not approach zero (minimum = 0.70, maximum = 0.97) (Tabachnick & Fidell, 1983). Additionally, eigenvalues did not differ greatly (maximum = .02, minimum = near zero) (Belsley, et al., 1980). The distribution of scores did not fit a perfect normal curve with a large cluster of scores toward the positive end of the distribution. This may again reflect an overuse of mid to upper gradings. The residual plots also showed evidence of minimal outliers. Reconsideration of these distributional problems did not alter conclusions regarding the non-significant outcomes. Nonparametric significance tests yielded similar outcomes regarding the view of linear relations between all measures (see Table 17),

Table 17

Spearman's Rho Between Measures Employed in Supplementary Set Two of the Military Sample

Scale	1	2	3	4	5
1. OAR	(.95)[a]				
2. SSE	.13	(.80)			
3. GSE	-.03	.51**	(.73)		
4. SM	.25	.29	.03	(.62)	
5. TKIM	.41**	-.00	.16	-.04	(.73)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to two items that made up the overall rating. The reader is cautioned that only the correlations between GSE and SSE and between TKIM and OAR manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

and inspections of scatterplots revealed no reason to suspect curvilinear outcomes. The considerations for correcting bivariate correlations for attenuation due to unreliability (Schmidt & Hunter, 1996, p. 201) between the presumed predictors and OARs were the same as in Set One. The exception to this was the revised version of the TKIM, which yielded the following corrected relationship; TKIM and OAR (r = .50, $p < .01$).

Organisational Sample

The data for the 87 respondents in Study One were imputed for missing data using EM. Data were imputed for 17% of assessment ratings. This reflected that of the OARs, 17% were not completed by the assessor group. Of the remaining measures, missing data were evident for 1% of self-monitoring ratings, 5% of tacit knowledge ratings and 0% for both specific self-efficacy, and general self-efficacy ratings.

Response rates from the individuals who participated in the AC were relatively low, and the human resource department of the department store under study reported that 429 individuals participated in the assessment process. Eighty-seven individuals, however, opted to participate in the present study, comprising a fairly low percentage of participation at approximately 20%. This may have been influenced to some degree by the length of the questionnaires in the study, particularly the TKIM. Caution must therefore be exercised with respect to non-response bias considerations in the present study.

An a priori power analysis was performed for multiple regression analyses with four predictors. This analysis revealed that the number of cases in this study was near the 85 cases necessary for a 2-tailed test at the .05 level of significance, at a

power level of .80 for medium effect sizes. The current analysis therefore achieved acceptable power, contingent on obtaining medium level effect sizes.

To investigate the possibility that those who did respond were a self-selected sample, and were not typical of the group as a whole, a *z* statistic was calculated to determine whether there was any difference between OARs of the sample individuals who participated in Study One, and those of the entire population from which the sample was drawn. The aggregated mean scores were provided by the company under study to eliminate any issues associated with anonymity. A 2-tailed *z* test failed to reject the null hypothesis that the sample and population means were equivalent $z(87, 429) = .07$, *ns*. This provides some evidence, with respect to OARs, that non-response bias was not an issue. However, there was no possible control, in this regard, for the measures of self-efficacy, self monitoring and tacit knowledge that were assessed in Study One, which may have been afflicted by non-response problems.

Table 18 shows the means and standard deviations for the measures used in the study. Particularly with respect to the OARs, small standard deviations could represent range restriction problems, as a small amount of variance in scores may not allow much opportunity for correlations to manifest. Table 19 shows the bi-variate correlations and internal consistencies for the same measures. All of the internal consistency coefficients were within the limits suggested by Nunnally and Bernstein (1994, p. 252). Significant bivariate correlations were found within the variables. In particular, positive correlations were found between general self-efficacy and specific self-efficacy ($r = .53, p < .01$). This relationship was the strongest with respect to overall magnitude, which could easily be expected of two measures of

Table 18

Overall Means and Standard Deviations for Measures Employed in the Organisational Sample

Scale	*M*	*SD*
OAR	3.08	0.45
SSE	6.20	1.04
GSE	6.18	1.34
SM	4.92	1.16
TKIM	4.06	1.96

Table 19

Bivariate Correlations Between Measures Employed in the Organisational Sample

Scale	1	2	3	4	5
1. OAR	(.79)[a]				
2. SSE	-.02	(.84)			
3. GSE	.15	.53**	(.71)		
4. SM	-.15	.24*	.39**	(.81)	
5. TKIM	.05	-.09	.05	-.11	(.72)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to items that made up the overall rating. The reader is cautioned that only the correlations between GSE and SSE, and between SM and GSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

self-efficacy. Self-monitoring correlated with specific self-efficacy ($r = .24$, $p < .05$) and general self-efficacy ($r = .39$, $p < .01$). The reader is cautioned that only the correlations

between GSE and SSE, and between SM and GSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

None of the predictors approached the conventional limits of significance when viewing the correlations between the set of presumed predictors and the DV, OAR. The overall pattern of the presumed predictors however, suggests the possibility of confounding, and therefore the need for multivariate analysis. The current research question aimed to investigate the relationship (if any) between the set of presumed predictors and OARs. The directionality of this relationship was ostensibly controlled in a temporal manner, by having participants complete questionnaires prior to the assessment process. The theory and research question did not suggest nor assume any causal relations outside of this temporal ordering. Neither did the study aim to consider subsets of variables separately. Given this, and the restrictive sample size of the study, standard all-in regression was selected as the most appropriate method by which to investigate these relationships.

The adjusted R^2 in Table 20 indicated that the predictors in the model above theoretically accounted for 4% (*ns*) of the variance in OARs in the population. The shrunken R^2 suggested that in different samples, the set of predictors would account for .04% (*ns*) of the variance in OARs. Note that a post-hoc power analysis revealed for an $R^2 = 0.085$, a sample size of 134 would be needed to achieve power of .80 with

Table 20

Multiple Regression Analysis for the Prediction of OARs in the Organisational Sample

		Partial Regression Weights	
Predictors	Raw (B)	Standardised Beta	95% Confidence intervals
SSE	-0.01	-.13	-.04 < B < .01
GSE	0.02	.31*	.00 < B < .03
SM	-0.01	-.24*	-.03 < B < -.00
TKIM	-0.00	-.02	-.01 < B < .01
Intercept	3.06		

Summary: R = .291(*ns*), R^2 = .085, Adjusted R^2 = .0404, Shrunken R^2 = .0004

* p < .05 (2-tailed)

four predictors in a regression model. In the case of this study, power was equal to 0.58, and thus the probability of making a type II error was 0.42 (Rosenthal & Rosnow, 1991). The effect found in this study was non-significant, although this study was afflicted with low statistical power associated with small sample sizes. The standardised partial regression weights suggest that GSE was the strongest predictor when applying this combination of measures, followed by SM, contrary to expectations, in a negative direction (although the sign of beta coefficients, of course, can be influenced by the modeler's choice of predictor combinations). None of the other predictors in the model displayed significant relationships with the criterion.

A residual analysis revealed no clear threat to homoscedesticity assumptions. Evidence assuaging multicollinerarity concerns was found with VIF indices being less

than 10 (maximum = 1.58, minimum = 1.04) (Chatterjee, et al., 2000), and tolerance indices did not approach zero (minimum = 0.64, maximum = 0.96) (Tabachnick & Fidell, 1983). Additionally, eigenvalues did not differ substantially (maximum = 0.01, minimum = near zero) (Belsley, et al., 1980). The scores were not normally distributed, with residual plots displaying a large clustering of scores toward the centre of the distribution. This may have reflected an overuse of central gradings. However, the significance tests used in multiple regression analyses are reasonably robust against violations of the normality assumption (Bobko, 2001). Reconsideration of these distributional problems did not alter conclusions regarding the non-significant outcomes. Nonparametric significance tests yielded similar outcomes regarding the view of linear relations between all measures (see Table 21), and inspections of

Table 21

Spearman's Rho Between Measures Employed in the Organisational Sample

Scale	1	2	3	4	5
1. OAR	(.79) [a]				
2. SSE	.02	(.84)			
3. GSE	.23*	.47**	(.71)		
4. SM	-.08	.27*	.28**	(.81)	
5. TKIM	.08	-.11	.06	-.05	(.72)

* $p < .05$; ** $p < .01$ (2-tailed)

Cronbach's alpha is provided in parentheses. [a]While Cronbach's alpha is not a measure of inter-rater reliability, it can be used as an estimate of intra-rater reliability (Schmidt & Hunter, 1996). Information on the specific allocation of raters was not made available. The alpha provided for OAR reflects internal consistency with respect to items that made up the overall rating. The reader is cautioned that only the correlation between GSE and SSE manifests as a statistically significant outcome after the appropriate Bonferroni adjustments are applied, so as to maintain study-wise type I error risk (at $p < .05$).

scatterplots revealed no reason to suspect curvilinear outcomes.

The bivariate correlations between the presumed predictors and OARs, corrected for attenuation due to unreliability (Schmidt & Hunter, 1996, p. 201) were as follows. SSE and OAR (r = -.02, *ns*); GSE and OAR (r = .02, *ns*); SM and OAR (r = -.19, *ns*); TKIM and OAR (r = .07, *ns*). These were considered to be similar to those correlations reported in the uncorrected bivariate correlations, thus no further correctional analyses were conducted. As previously mentioned, the corrected coefficients may have been higher than reported here, had a proper index of interrater reliability been available for OARs (Schmidt & Hunter, 1996, p. 209).

Discussion

Study One sought to investigate the extent to which a specific set of traits that were not formally assessed in the AC context explained meaningful variance in OARs. The set of constructs included domain specific (SSE) and general self-efficacy (GSE), self-monitoring (SM) and managerial tacit-knowledge (TKIM), as suggested by Klimoski and Brickner (1987), Chan (1996) and Arthur et al. (2000). Overall, the results of Study One showed differential outcomes across the two samples employed in the study. In Set Two of the military sample, the full combination of these constructs explained slightly more variance in OARs in the population than in Set One of the military sample and the organisational sample. However, the only significant contributor in the model was tacit-knowledge. Furthermore, its manifesting measure in this analysis, the TKIM, exhibited poor internal consistency in this sample. Also, when correcting for validity shrinkage when generalising across samples, the combination of the constructs measured in Set Two explained very little variance in

OARs. In the organisational sample, the combination of these constructs explained very little variance in OARs.

Military Sample

Due to practical difficulties with respect to the TKIM, the military sample needed to be split into two sub-sets. Set One examined the full sample of 100 participants, and the relationship between the predictors SSE, GSE and SM with the criterion, OARs. At the bivariate level, no correlations of any magnitude or significance were seen between the set of predictors and OARs. The same was evident at the multivariate level with no significant contributions from individual predictors. The composite of the predictors explained around 2% of the variance in OARs in the population. When correcting for validity shrinkage when generalising across samples, this figure dropped to .02% of the variance explained in OARs. Thus, it is unlikely that these constructs had much bearing on the prediction of OARs in Set One of the military sample.

Set Two of the military sample examined 44 participants from the total sample pool. At the bivariate level, the results were different from Set One, in that significant individual correlations of comparatively moderate magnitude were found between SM and TKIM with OARs. Note that these relationships were non-significant when the appropriate Bonferroni adjustments were considered. At the multivariate level, however, only the TKIM remained significantly related as an individual predictor of OARs. The composite R^2 in Set Two increased to around 16% ($p < .05$) of the variance in OARs in the population. The amount of variance explained in OARs by this composite dropped to around 1% when correcting for validity shrinkage across samples. A supplementary analysis on these data, which corrected the TKIM for

unreliability, found that the composite explained more variance in OARs in the population ($R^2 = 23\%, p < .05$). This figure dropped to around 2% of the variance explained when generalising across samples.

In consideration of these findings, it could be that sample characteristics may determine the extent to which these latent constructs will manifest and/or influence assessments in these contexts. In the organisational sample, the position being assessed was at an entry level, with little opportunity for managerial type behaviours and promotion. However, in the military sample, the positions being assessed had potential for promotion and the elicitation of managerial type behaviours. Thus, assessors may have been more likely to pick up on these tacit managerial behaviours, and likewise, the candidates may have been more prone to impart such behaviours. Considering that managerial tacit knowledge was not a construct that was formally assessed in this AC, 16% of the variance explained by an external measure of the construct could be construed as fairly sizable, or at least notable and considerable. This may, however, not be true when generalising across samples, as the shrunken R^2 for this sample suggests.

Organisational Sample

In the organisational sample, at the bivariate level, no correlations of any magnitude or significance were found between OARs and the individual variables included in the analysis. The outcome of the multivariate analysis was a low and non-significant R^2, for which the composite explained 4% of the population and .04% of the sample-generalisable variance in OARs. Two of the individual predictors were significant in this model, namely GSE and SM. While GSE was significant in the expected direction, SM was unexpectedly negative in its relationship with OARs.

Thus, from the organisational sample, the results suggest that the individual and composite contributions of SSE, GSE, SM and TKIM are probably unrelated to performance in the AC employed in this sample. That is, counter to the possible explanations for the construct validity of ACs presented by Klimoski and Brickner (1987), Chan (1996) and Arthur et al. (2000), it would appear that managerial assessors were not tapping into latent constructs that were informally assessed during this AC. Also, counter to Moser, et al., (1996) a significant and negative relationship was found between SM and OARs. However, given the small amount of variance explained by the composite of these predictors, and evidence for the presence of multiple outliers, the negative relationship between SM and OAR was probably spurious. The substantial differences seen when comparing standardised betas across otherwise similar models (tested in these two samples) are disconcerting. But it also brings to mind the potentially powerful influence on these betas that can be caused by otherwise 'model-irrelevant' sample differences (Rosenthal & Rosnow, 1991).

Considerations

The samples in Study One were hampered by low power, as post-hoc power analyses revealed that these studies tended to stack the odds in favour of the null hypothesis and limited the analytical sophistication applied. This problem is, unfortunately, common in the psychological literature (Schmidt, 1996). The organisational sample suffered from a large degree of non-response to questionnaires. Thus, those who did respond may have been different in some way from those who did not, and all findings related to this sample should be interpreted with this in mind. The degree of non-response shown was possibly influenced to some extent by the length of the TKIM, which is a time-consuming questionnaire. Researchers interested

in such notions as managerial tacit knowledge as measured by questionnaires should possibly develop shorter versions, or find alternative methods.

The TKIM also exhibited low internal consistency in one of the samples. This may have been influenced to some degree by sample characteristics (e.g., respondent acquiescence), or it could be that more work needs to be done to refine the measure so that it holds greater coherence with regard to its measurement, and so that it is more manageable time-wise. Note that the sample had to be split because the first group of participants were issued a version of the TKIM that required a group of subject matter experts to create profiles from which deviation scores could be calculated. Non-response from the subject matter experts was such that the first administration of the TKIM had to be abandoned. Again, this was possibly related to the TKIM's length.

Analytical Limitations

Had the organisations been able to provide a much larger sample of participants, these analyses could have included the development of a measurement model, and a structural path model via covariance structure modelling (e.g., LISREL, AMOS, or SAS PROC CALIS applications). This would have allowed for more detailed considerations of test item performance in this context, and examination of modelling of correlated error terms. The latter, for instance, could manifest due to self-presentation, self-deception motives (Paulhus, 2002) or aspects of self-efficacy. This, in turn, could be evident in correlated error terms for GSE and SSE items targeting over-lapping self-efficacy content domains. These, in turn, could inflate the apparent correlations between GSE and SSE in the simple analyses herein (see Tables 8, 12, 15, and 19). The sample size needed for these more sophisticated models could have assumedly been several times that provided by these organisations.

Theoretical Implications

In the Organisational Sample and Set One of the Military Sample, very little variance was explained by the composite of the suggested group of latent variables that were hypothesised to explain a meaningful amount of variance in OARs. It would seem, therefore, that if the pre-AC self-reports in this sample are to be believed, then something may be amiss (e.g., latent constructs influencing OAR variance or that latent trait measurement on the basis of the traits included for analysis did not offer an acceptable representation of what was actually measured in the AC). Thus, given the enduring exercise effect that occurs in ACs (see Hough & Oswald, 2000, for a review), one could speculate that in this sample, other variables were associated with OAR variance. Chan (1996) suggested that in order to predict stable criteria, an AC must tap stable trait-based variables. Thus, perhaps the choice of underlying constructs in this study needs revision, and perhaps further examination of such variables as intelligence and general mental ability, which have already been found to relate to OARs (Klimoski & Brickner, 1987; Schmidt & Hunter, 1998) as well as exhibitionism (Fleenor, 1996), conscientiousness, extraversion (Furnham, et al., 1997), and social confidence (Moser, et al., 1996) is necessary. This said, the set of constructs in the present study were conceptually similar or related to many of the constructs above. Greater consideration should also be given to how assessors actually conceptualise assessee performance related to these criteria and how these assessor conceptualisations influence OARs.

It could also be that the latent concepts in the present study were not measured in a satisfactory manner. In rudimentary terms, it has been suggested that the type of person who might get ahead in ACs will display a certain degree of self-efficacy

about their performance and themselves, that they will skilfully be able to present a positive impression on others, and that they will be savvy with regard to what it takes to be a successful manager (Chan, 1996; Klimoski & Brickner , 1987). The present study sought to gain an indication of these concepts through a questionnaire based format. However, perhaps such a format is not conducive to measuring the holistic sum of such socially contextualised variables. It may even be that the influence that an individual has over a group determines, to some degree, the success of an individual in an AC. As ACs commonly measure behaviour whilst in a group, it is possibly difficult to separate these characteristics from their group context. Thus, these constructs might be better judged formally in a group context by a panel of observers. Indeed, such variables theoretically depend, to some degree, on the opinion of the 'others' (Sartre, 1964). Specifically, it is up to the assessor group to make interpretations of the behaviours described. These exerted behaviours may differ from the self-beliefs of the individual. Indeed, Fletcher and Kerslake (1992) believe their evidence confirmed that self-assessments are not always accurate in AC contexts.

Set Two of the military sample did explain a comparatively sizable amount of the variance in OARs through the composite of the predictor variables. This was most clearly attributable to the TKIM, a finding that is indistinct to some degree because of the lack of internal consistency associated with the TKIM measure in the military sample. It may be, as previously argued, that certain latent traits are important in some contexts and not others, depending on the demands of the job and the characteristics that the assessors endeavour to detect.

A shift from Chan's trait-predicts-trait suggestion is the notion that a different paradigm might explain the predictive utility of ACs. Indeed, the trait paradigm is not

the only theory that might yield predictive validity. For instance, Hunter and Hunter (1984) and Schmidt and Hunter (1998) found work samples to be the most highly predictive form of selection in their meta-analysis of selection methods. Rather than applying trait theory, work samples operate on the simpler notions of behavioural consistency theory (Cook, 1998), that is, like behaviour predicts like behaviour. An exploration of possible alternative paradigms explaining the predictive utility of ACs has been suggested by very few researchers (Herriot, 1986; Lowry, 1997; Robertson, et al., 1987). Thus, against the argument for latent trait measurement, it may be that the AC ratings are being used directly by non-psychologist assessing groups, but in a different way than that supposed under trait theory.

Explanations of trait measurement in terms of traits that are not formally assessed in ACs could shroud the meaning behind OARs when they are conceptually unrelated to formally measured traits. It could be argued that such divergent assessment is undesirable, because with arbitrary and unintended measurement, it is possible that a certain degree of precision will be lost. It may even call into question the use of and expense associated with ACs (Schmidt & Hunter, 1998). A lack of measurement clarity is likely to affect the quality of decisions that are made on the basis of AC ratings. Although there are studies that have found relationships with OARs and variables that are similar in some respects to those conceptualised in the present study (Furnham, et al., 1997; Fleenor, 1996; Moser, et al., 1996; Schmidt & Hunter, 1998), the findings (in the present study) add some evidence to those researchers, such as Goffin, et al. (1996) and Chan (1996), who found no relationship between externally measured traits and OARs. The exception to this (in the present study) was, perhaps, tacit knowledge within a particular sample.

Chapter Three: Study Two, Perceptions of Assessors and Candidates with respect to Measurement Models

Method

Candidates

To address the research questions raised in the hypothesis section, the military sample described in *Study One* was administered a short questionnaire. Two versions of the questionnaire were developed: one for candidates, the other for assessors. Data were collected from a selection board, the military equivalent of an AC, that was already in existence and was used for recruitment and selection purposes in the Royal New Zealand Air Force (RNZAF) in Auckland, New Zealand. Data were collected at two periods during the year from the 31st of July, 2000 to the 18th of August, 2000, and again from the 2nd of October, 2001 to the 22nd of October, 2001. Questionnaire responses were collected from 100 potential recruits. Demographic information for this sample is presented in Table 22.

Assessors

Assessing participants included 27 male officer-level personnel from the RNZAF located in Auckland, New Zealand. Demographic information for this sample is presented in Table 23. Ethnicity could not be assessed due to confidentiality concerns expressed by the organisation under scrutiny. All assessors had previous experience in assessing participants in multiple selection boards, although none had received any post-

Table 22

Demographic Statistics, Candidates, Study Two Military Sample

N = 100	Frequency (= %)
Gender	
Male	83
Female	12
Non Responders	5
Ethnicity	
Caucasian	75
Asian	3
Maori and Pacific Islanders	16
Other	2
Non Responders	4
Age	
15-20	44
21-25	33
26-30	7
31-35	4
Non Responders	12
Education	
No formal education	1
School Certificate	3
Sixth Form Certificate	36
Bursary	32
Bachelor's Degree	14
Higher University Degrees	2
Other	3

Table 23

Demographic Statistics, Assessors, Study Two Military Sample

N = 27	Frequency
Gender	
Male	27
Female	0
Age	
21-25	6
26-30	6
31-35	8
36-40	2
41-45	4
51-55	1
Education	
Bursary	4
Bachelor's Degree	4
Masters Degree	10
Other	7
Non Responders	2

high school training in psychology. All assessors had at least two years experience and were regarded as subject matter experts of the position being assessed.

Measures: Candidates

The candidate version of the questionnaire included 3 sets of questions, which are presented individually in the results section. The various models presented several different possibilities for methods of assessment. The first and second set of questions solicited information on the model under which the candidates perceived that they were

being assessed. The models presented in the first set of items, ranged from the attribution of stable characteristics to the identification of exercise specific behaviours (see results section). Candidates were required to rate the extent to which they perceived they were assessed under a given model. Ratings were scored on a 5-point scale ranging from 1 (never assessed under); 2 (seldom assessed under); 3 (unsure); 4 (often assessed under); to 5 (always assessed under).

For the second set of items the candidates were required to specify the extent to which they perceived that certain models guided the assessment of behaviour. Again, this was rated on a 5-point scale ranging from 1 (extremely irrelevant); 2 (irrelevant); 3 (neither irrelevant nor relevant); 4 (relevant) to 5 (extremely relevant). Prior to the third set of questions, the candidates were informed that there are often problems associated with the measurement of stable attributes in AC related evaluation procedures. Given this information, the final set of questions again asked the candidates to rate which model they assumed they were being assessed under on a 5-point scale ranging from 1 (never used); 2 (seldom used); 3 (unsure); 4 (sometimes used) to 5 (always used). Means and standard deviations were calculated for each item.

Measures: Assessors

The assessor version of the questionnaire was similar to the candidate version, with an additional section, aiming to tap into the extent to which the assessors found utility in, and saw evidence of, the specific dimensions assessed in the evaluation process. Concerning the use of the term 'model' in this study, it was expected that the assessor group would have some understanding of the meaning of the term, because the Air Force

stated that they provided training which covered such issues. Regardless of a particular assessor's potential comprehension of the term 'model', exactly what was meant by each of the models presented was clearly stated. The various models presented several different possibilities for methods of assessment. There was no competition between items, in that candidates could potentially rate any or all of the models high or low, depending on their perceptions.

The first set of items enquired as to whether the assessors perceived that each dimension was useful for evaluating and distinguishing between candidates. This was rated on a 5-point scale ranging from 1 (extremely inadequate); 2 (inadequate); 3 (neither inadequate nor useful); 4 (useful) to 5 (extremely useful). The second set of items sought to gauge which ability traits were seen as being exhibited by candidates across all of the exercises during the selection board. The ratings ranged from 1 (never shown); 2 (seldom shown); 3 (unsure); 4 (often shown) to 5 (always shown).

The third set of items sought to gain insight into the relevance of various models of assessment during the selection board. Models presented ranged from the attribution of stable characteristics to the identification of exercise specific behaviours. The ratings ranged from 1 (extremely irrelevant); 2 (irrelevant); 3 (neither irrelevant nor relevant); 4 (relevant) to 5 (extremely relevant). As in the candidate version, the assessors were, at this stage, informed of some of the problems associated with attempting to measure stable characteristics in AC related processes. The purpose of this step was to see how such knowledge might influence the assessor's judgment of the models under which they were assessing. Given this information, a range of models were presented, and candidates were asked to rate their perceptions as to the extent to which they utilised the respective

models when assessing. Ratings ranged from 1 (never used); 2 (seldom used); 3 (unsure); 4 (sometimes used) to 5 (always used).

Results

In this chapter, the perceptions of candidates and assessors were explored. The study focused on the models under which both candidates and assessors perceived they were being measured, and under which they perceived were measuring behaviour, respectively. Due to the exploratory nature of this research, and because of the nature of the questions asked, the analysis in this chapter was purely descriptive. This included the calculation of the mean and the standard deviations to responses on each item. Overall results suggested that regardless of the model presented, both candidates and assessors tended to perceive that all of the models for assessing behaviour presented to them were relevant or useful to some degree in the AC context.

Candidates

In this study, there were no missing values in the data obtained from the same military sample as in Study One, possibly due to the conciseness of the questionnaire. The sample size was slightly larger than Study One in terms of overall respondents to this particular questionnaire (N = 107). The population of applicants totalled 116, thus a response rate of approximately 92% was considered acceptable.

Both Tables 24 and 25 show the mean responses and their respective standard deviations for each of the models presented, that could potentially be assumed to underlie and guide the assessment during the selection board. All of the mean ratings for each model presented appeared to vacillate slightly around the 4th point. This reflected, in Table 24, that candidates were inclined towards the rating signifying that they were often assessed under all of the respective models (a rating of 4). In Table 25, the pattern of responses was similar, and regardless of the model, the responses appeared to fluctuate slightly around the 4th point, reflecting that regardless of the model presented, candidates tended to rate overall, that the model was relevant (a rating of 4).

Table 24

Model Assumed to Underlie Assessment: Candidates

Statement	Mean Rating	*SD*
Ability traits that were stable across different exercises	4.09	0.77
Ability traits within individual exercises only	3.65	0.91
Behaviour that was stable across the different exercises	4.17	0.84
Behaviour within an individual exercise	3.93	0.90

Table 25

Model Assumed to Guide Assessment: Candidates

Statement	Mean Rating	*SD*
The identification of traits that were relevant within individual exercises	4.02	0.50
The identification of traits that were stable across different exercises	4.24	0.64
The actual behaviour of candidates in individual exercises	4.12	0.75
The actual behaviour of candidates that was stable across the different exercises	4.33	0.63

At this stage, candidates were informed that research had shown that ability traits measured via AC based methodology often did not appear to have been measured in the way that they were intended. Given this information, candidates were asked to give their perception of the model that they thought they were assessed under. The rationale here revolved around the possibility that the candidates might exhibit demand characteristics in the presence of this information.

Again, similar patterns of acquiescent responses were exhibited in Table 26. Despite being warned of the possibility of measurement problems associated with trait-based judgements in ACs, candidates continued to rate towards the 4th point on the rating scale, which tended to reflect, in this case, that they perceived each of these models were sometimes used by the assessors.

Table 26

Model Assumed to Guide Assessment After Being Informed of the Measurement Problems in ACs: Candidates

Statement	Mean Rating	*SD*
Ability traits that were stable across all exercises	4.18	0.75
Ability traits that were specific to individual exercises	4.05	0.75
The behaviour of candidates that was stable across all exercises	4.24	0.80
The behaviour of a candidate rated on individual exercises	4.01	0.80

Assessors

Data were imputed for the responses of the 27 assessors in the military selection board. Roughly 2% responses were missing, which constituted an acceptable response rate to the questionnaire items of roughly 98%. Again, for assessors, it was decided, due to the nature of the questionnaire and the exploratory nature of the questions under scrutiny, that a descriptive analysis would be utilised.

Table 27 shows the averages and standard deviations with regard to perceptions of the usefulness of individual dimensions. All of these perceptions vacillated around the 4th point in the scale, which reflected that overall, the assessors found the set of dimensions to be useful for evaluating and distinguishing between

Table 27

Usefulness of Individual Dimensions: Assessors

Dimension	Mean Rating	*SD*
Written Communication	3.08	0.92
Oral Communication	4.33	0.48
Stability Under Pressure	4.74	0.45
Relations With Others	4.37	0.79
Group Influence/Leadership	4.85	0.46
Initiative	4.30	0.67
Determination	4.22	0.58
Reasoning/Planning	4.44	0.64
Decision Making	4.44	0.51

candidates. The exception to this was written communication, for which the average rating tended towards the 3rd point in the scale.

Table 28 shows the extent to which the assessors perceived that they saw evidence of particular dimensions exhibited across all exercises. Overall, the ratings tended towards the 4th point, which reflected that the dimension was perceived as often shown across the exercises. The exception to this was written communication, which was rated lower overall and tended toward the 2nd point. This reflected that the dimension was perceived as seldom shown overall.

Table 28

Dimensions Perceived as Being Seen Exhibited Across All Exercises: Assessors

Dimension	Mean Rating	*SD*
Written Communication	2.04	0.34
Oral Communication	4.48	0.51
Stability Under Pressure	4.07	0.62
Relations With Others	4.41	0.57
Group Influence/Leadership	4.19	0.48
Initiative	4.15	0.53
Determination	4.19	0.74
Reasoning/Planning	4.15	0.77
Decision Making	3.93	0.83

Table 29 shows the extent to which the assessors perceived that certain models were assumed to guide the assessment of candidates. Assessor' ratings tended toward the 4th point on all models presented. Table 30 shows that this effect endured information regarding the measurement problems associated with AC based evaluation, and again, regardless of the model presented, the assessors tended toward the 4th rating. As with the candidate group, information on measurement problems was presented to estimate the effect of demand characteristics. This suggested that they perceived, overall, that all of the models were at least sometimes used by the assessor group when rating candidates.

Table 29

Model Assumed to Guide Assessment: Assessors

Statement	Mean Rating	*SD*
The identification of traits that were relevant within individual exercises	4.49	0.61
The identification of traits that were stable across different exercises	4.50	0.56
The actual behaviour of candidates in individual exercises	4.57	0.50
The actual behaviour of candidates that was stable across the different exercises	4.44	0.63

Table 30

Model Assumed to Guide Assessment After Being Informed of the Measurement Problems in ACs: Assessors

Statement	Mean Rating	*SD*
Ability traits that were stable across all exercises	4.25	0.70
Ability traits that were specific to individual exercises	4.60	0.61
The behaviour of candidates that was stable across all exercises	4.22	0.68
The behaviour of a candidate rated on individual exercises	4.50	0.67

Discussion

Given the multitude of research that has a found a lack of construct validity in AC ratings (Arvey & Murphy, 1998; Schmidt & Ones, 1992), this exploratory study sought to gain a preliminary insight into the extent to which assessors and candidates in ACs perceived that they were being assessed under alternative and differing measurement models. Overall, the results of Study Two showed that when assessors and candidates were presented with a range of different models for assessment, they tended to perceive that all of the models were relevant or useful to some degree in the AC context.

Candidates

This finding is of little concern with respect to the candidates who participated in the AC itself. One could easily expect candidates to have little idea as to the assessment model that they were assessed under. Indeed, the pattern of acquiescent responses seen in Study Two reflects this with, overall, all of the different models being rated in congruence with the notion that the particular models were 'relevant'. This finding draws Kleinman's (1993) dimensional transparency suggestions into consideration. In the AC under scrutiny, the dimensions that candidates were being assessed under were not revealed to the candidates at any time. Such information was kept confidential by the RNZAF. Had the dimensions been made explicit to the candidates, they may have been made more aware of the assessment model in practice. Such revelations are questionable, however, in terms of how they might improve the process in terms of its ability to detect individual differences. When candidates are informed of the assessment dimensions, one might question the extent to which an assessment, under these circumstances, reflects an individual's true performance. There remains the possibility that such actions could increase the likelihood that impression management behaviours are manifest.

Assessors

Of more concern is the finding that the assessors did not, overall, distinguish between the different models of assessment they were presented with. It was possible that the assessors found the wording of the items in the questionnaire confusing and difficult to comprehend. Not only was this evidenced in the acquiescent responses that were similar in overall trend to the responses of the candidates, but it was also evidenced in comments that were written on many of the questionnaires (to the effect

that the assessors found the wording of item statements confusing). Whilst it is reasonable that candidates should not be able to make such differentiations, it is without question that when assessors are using trait-based models they should understand what a trait is, and what the underlying notions are that act as the basis for traits. This includes an understanding of the terminology behind such models.

If organisations intend to utilise models that are founded on trait-based assumptions, then assessors should understand the difference between a situationally specific behaviour, and a relatively stable and enduring characteristic. If the former is believed, then the exercise effect will, in all probability, prevail. If the latter is believed, then it is more likely that stable characteristics will be measured. If both models are being utilised, then assessors may assess in terms of a mixture of traits and situationally specific behaviours. This was found to be the most common model across 34 studies (Lievens & Conway, 2001). This mixed model possibly has awkward pragmatic implications, particularly for feedback and selection, as discussed in the introduction. Fundamentally, it is the intention to capitalise on and utilise the assumptions of the trait paradigm in dimension-specific ACs. When decision makers use the results of this model, they assume measurement precision. In this regard, I/O psychology needs to improve the tools that are provided to decision makers.

In the case of the present study, it was generally found that assessors thought all models presented to them were relevant to the assessment. Such a failure to distinguish between the models could be the result of a lack of training on such factors. The selection board under scrutiny, however, was one of the most respected AC processes in New Zealand, which utilised the knowledge of highly trained military officers to make employment decisions pertaining to potential candidates. Next to training there is another factor that requires consideration. There remains the

possibility that to assess traits, an assessor needs to be trained as a psychologist, as evidenced in both an implied and salient sense across several studies (Arthur et al., 2000; Gaugler et al., 1987; Lievens & Conway, 2001; Sagie & Magnezy, 1997). That is, it may take several years of training before the notion of trait measurement is clearly understood and is of use to the assessing personnel.

Considerations

The present study is possibly limited to some extent by the complexity of the items in the questionnaire. As previously discussed, candidates would probably not be able to distinguish between measurement models in any case unless, perhaps, they had training in psychology. A repeat of this study could possibly benefit from items of a simpler nature for a candidate/assessor group, or even a pictorial description of the concepts to ensure clarity for the participants. It could be argued, however, that assessors should be well aware of the distinction between these concepts and the related terminology well in advance of ever using these concepts to make employment decisions. Also with respect to the scale used in the first set of questionnaire items for assessors, a repeat of this study would most likely benefit from a scale that used the same continuum for scale anchors. In the present study, the scale ranged from extremely inadequate to extremely useful, whereas it should have ranged from extremely inadequate to extremely adequate or extremely useless to extremely useful.

The generality of the findings in this study is limited by the restrictive sample size, particularly for the assessor group at 27 participants. However, the assessors in this study are, in all probability, more highly trained and/or qualified in terms of expert knowledge related to the target job than assessors employed by many consulting firms in practice (Fletcher & Anderson, 1998). Additionally, a group of 27

assessors is relatively sizeable, considering that most ACs in one run will typically use only three assessors (Ballantyne & Povah, 1995). The generalizability of Study Two is, however, restricted to a military sample. Such samples may be sufficiently unique organisationally or culturally that further research is necessary to generalise this elsewhere.

Consideration must also be given to the exploratory nature of this research. As neither any particular directionality nor outcome was expected, the research was designed to be descriptive. As a result, only means and standard deviations (SDs) were calculated to reflect overall responses to particular items. Thus, a level of detail was lost through the aggregation of ratings. However, the SDs were relatively small considering a 5-point scale was used, and histograms of item responses typically showed a clustering of responses toward the 4th and 5th points in the scale.

Theoretical Implications

The results of Study Two highlight a potential issue, from an exploratory stance, that should be addressed in greater detail in future studies. That is, assessors should be audited with regard to their understanding of the models they use for assessment. The candidate group provided an interesting comparison in this regard. Both assessor ratings and candidate ratings tended towards agreement that all of the different measurement models presented were useful in the appraisal of AC candidates. This could create confusion in terms of the criteria under which ensuing decisions need to be made. Future studies on ACs based on the trait paradigm should look at methods of training that will help to improve assessor perceptions in this regard. Such studies should also look to see if such an understanding would lead to more construct valid ratings.

A less complicated set of items, or a qualitative study might assist to understand the degree to which non-psychologist assessors suffer a lack of comprehension about the trait paradigm and its use in an AC. If such a case were true, it might be that an alternative, less complex, paradigm would be more appropriate. The obvious example relates to the behavioural paradigm, which takes a more simplistic view of human behaviour that may be easier for non-psychologists to comprehend.

Chapter Four: Study Three, A Comparison of Task-Specific and Dimension-Specific ACs

Method

Participants

Data were collected from an AC that was constructed for a large private sector retail chain named Farmers Trading Company based nationwide throughout New Zealand. The AC was used approximately 42 times from the 20th of September until the 7th of October, 2002, for the selection of retail and customer service workers performing the specific functions of general sales people. The main functions of this position were to tend tills and sell products directly to customers. Two hundred and forty participants comprising 44 males and 168 females with a mean age of 31.274 years (SD = 12.318) were assessed. Twenty-eight candidates did not respond to the age item. Demographic information for this sample is presented in Table 31.

Assessors

Assessing participants were 11 managerial staff members from the retail organisation described above. Only six of the assessors completed the demographic questionnaire. Of those who completed the questionnaire, there were three males and three females, with a mean age of 33.5 (SD = 10.932) located in Auckland, New Zealand. All assessors described themselves as Caucasian. Educational demographics were represented by one assessor who described holding no formal education, two who

A pilot for Study Three is presented in Appendix I

Table 31

Demographic Statistics, Candidates, Study Three Private Sector Sample

N = 240	Frequency	%
Gender		
Male	44	18
Female	168	70
Non Responders	28	12
Ethnicity		
Caucasian	180	75
Asian	3	1
Maori	10	4
Polynesian	5	2
Indian	4	2
Other	11	5
Non Responders	27	11
Education		
No formal education	54	23
School Certificate	57	24
Sixth Form Certificate	40	17
Bursary	26	11
Bachelor's Degree	16	7
Master's Degree	1	0
Other	18	8
Non Responders	28	12

reported holding school certificate, and three who reported holding sixth form certificate. Due to the low response rate on the demographic questionnaire, the personnel department of the company under scrutiny was questioned to verify information about the missing responses. All assessors had previous experience in assessing potential recruits in

multiple ACs, although none had previously received either FOR training or any form of psychological training. All assessors had over two years experience in, and were regarded as subject matter experts of, the position being assessed.

The AC

After developing a policy statement detailing the purpose to which the AC would be put, and who would be involved in the process, the initial step in the construction of the AC was to execute a competency analysis of the target position. Again, the purpose of the current study was to compare a task-specific model with a dimension-specific model. As such, the competency analysis involved an improved repetition of the pilot for Study Three (see Appendix I), where a two-tiered process was employed producing a detailed task-analysis (gathering information on the tasks that organisational members perform), and then a classification and, moreover, an extrapolation of these tasks into dimensions.

The job analysis approach in the present study differed from the pilot for Study Three in that a mixed deductive and inductive approach was taken for the analysis of the position. Inductive approaches seek to find new and specific information on a job, whilst deductive approaches start with an existing body of information about jobs, and the researcher subtracts information from that framework that is not relevant to the job under scrutiny (Peterson & Jeanneret, 1997). While the intention was to concentrate on assessing highly job-specific task and dimensional information, there was no agenda in existence for the particular information that should be assessed. One might expect such an agenda to arise in the context of a development centre (where there may be specific

developmental concerns and performance gaps), however, this notion is theoretically less likely to arise in the context of selection procedures, where the focus should be on the most precise and objective form of assessment available. The behavioural and trait-based methods of analysis in this study utilised an inductive task analysis for the behavioural or task-related aspects of the job. For the dimension or trait-based information, the current study employed a deductive approach known as Threshold Traits Analysis (TTA, Lopez, 1988).

Task Analysis

The first stage of the competency analysis involved utilising job analysis methods to identify the key tasks that were implicated in the successful performance of the general sales, one-on-one sales, and merchandiser positions. Note that although ACs were developed for all three of these positions, only the AC developed for the general sales role was studied. This is because the department store only needed to select for this role at the time this research was conducted. The task analysis involved a review of the job descriptions already in existence coupled with interviews incorporating incumbents and supervisors, comprising the panel of subject matter experts (SMEs). The human resources department of the present company had created comprehensive documentation through inductive SME interviews and direct observations that detailed specific task information on all of the roles under scrutiny.

The existing task-related documents formed the basis of interviews, which were held with nine incumbents for each position across three outlets of the company. That is, there were 27 SMEs in total, who were geographically dispersed throughout Auckland.

Some researchers suggest that only three SME's are needed for procedures such as these (Aamodt, 1999; Green & Stutzman, 1986), and in combination with the existing task-based information being less than six months old, it was felt that this SME sample would be sufficient for the present analysis. Additionally, there exists the possibility that incumbents might present a different account of their jobs than their supervisors and managers. The present study employed SMEs from both levels to counteract this possible confound. All information was presented to a panel of managerial SMEs for verification and critique. Additionally, biographical information pertaining to the task analysis respondents could not be obtained because the small numbers of analysts could easily lead an individual to feel they might be personally identifiable on the basis of such information. Biographical information pertaining to the respondents implicated in the task analysis was not deemed important for the replicability of the study, as the important criterion for replication was the six months experience of the analysts in their respective positions.

In an interview situation across three different outlets, all SMEs were requested to critique and discuss the existing task-related information, in particular, to relay whether they thought information had been omitted, was inaccurate or was superfluous to the actual requirements of the job. To give the analysis a strategic outlook, interviewees were asked to give their views on what tasks they thought might be important for their positions in a future context, as suggested by Thornton (1992) and Woodruffe (1993).

Guided by the course of action set out by Lowry (1997) for task-specific AC construction, a questionnaire was developed listing the tasks derived from the information obtained in the task analysis. The purpose of this questionnaire was to

determine the relative rank and importance of particular tasks. Three questions were asked of incumbent level SMEs with respect to each task, including: (a) a dichotomous dimension on whether the incumbent actually performed the task or not, (b) the criticality of a particular task relative to other tasks for successful operations, and (c) the importance of a particular task upon entry to the position, relative to other tasks. The last two dimensions were both rated on generalised 6-point scales ranging from 0 (this task was not performed in the position) to 5 (this task was extremely important to the position).

Item a) above differed from the relative time spent scale suggested by Lowry (1997). Lowry's original scale was replaced for three main reasons. Firstly, it was felt that a dichotomous dimension might save time for the already heavily cognitively burdened SME panel. Second, research suggests that many of the scales in task analyses appear to measure similar constructs (Sanchez & Frazer, 1992), and it was thought that adding a relative time spent scale would add little incremental information over and above criticality and importance related scales. The decision was also made to remove the relative time spent scale because of the argument that although some tasks might be performed rarely or irregularly, it does not necessarily follow that the task is not critical or important to the job (Harvey, 1991). As the major goal of the present task analysis was to identify those tasks that were most important, it was felt that the relative time spent dimension should be replaced in favour of the dichotomous dimension.

From the task level information obtained in the questionnaire detailed above, the most critical tasks were selected for inclusion in the AC exercises. In concurrence with the suggestions of SMEs, a checklist of the typical actions that would be required to successfully perform the selected set of tasks was developed. These typically involved

short checklists of around 8-15 actions considered important for the successful completion of a given AC exercise. As Lowry (1997) emphasised, no inference of the existence of complex constructs was made at this stage.

Six SMEs responded to the task analysis questionnaires concerning the Merchandiser and One-on-one positions. Three SMEs responded to the General Sales task analysis questionnaire. These response rates were not deemed problematic because the information obtained from the task analysis would be presented to a second panel of SMEs in managerial positions to increase the number of SMEs who had input into the task analysis (and also the trait-based competency analysis). This step was also employed to gain a more holistic perspective because incumbent and managerial levels might present slightly different perspectives with regard to the aspects of particular jobs.

With lower numbers of respondents, tasks were included where over 50% of the respondents agreed that a task was actually relevant to the position at all. This criterion was set reasonably low purposely, so as to allow for a greater number of tasks to be included in the final analysis for the managerial SME panel to scrutinize. The median scores for criticality and importance on entry to the position were calculated for each task. The median was chosen because as a measure of central tendency, because it is less likely than the mean to be affected by skewed distributions that might occur as a result of small subject numbers. These median scores were then multiplied by each other so that any task rated as 0 (meaning the task was not relevant) would be excluded and so that any task rated as 1 (meaning that the task was of little importance) would remain at a low rating, as suggested by Lowry (1997). These multiplied values were then rank ordered so

that the most important tasks appeared first. It was at this point that the competency analysis for the task-specific AC model concluded.

TTA (Threshold Traits Analysis)

As opposed to the inductive method of job analysis detailed above for the attainment of task level information, the job analysis for the identification of traits for the dimension-specific component of the AC was obtained through a deductive instrument known as the Threshold Traits Analysis (TTA) (Lopez, Kesselman & Lopez, 1981).

Permission to use the TTA was obtained from Lopez and Associates, a private consulting firm, based in New York. The latest version of the TTA presents a range of 36 characteristics, traits or dimensions that are thought to be important across a range of different jobs. These 36 characteristics cover five trait-based categories, comprising physical, mental, learned, motivational, and socially derived factors. When completing the TTA, respondents must firstly decide if the trait, as described in the questionnaire, is important to their position with a dichotomous yes/no response. If the response is negative, the trait is ignored from further analysis, hence the deductive nature of the TTA. For traits that are deemed relevant, the respondent then decides upon the magnitude of the trait, labelled trait-level, which would be required for clearly acceptable performance in the position. Lastly, the respondent must decide upon the magnitude of the trait that is required for superior performance in the position. These last two decisions are made on the basis of 4-point scales that reflect the increasing magnitude of a given trait, from level 0 to level 3.

TTA Respondents

To obtain acceptable reliability for the instrument, it is advised that no less than five independent TTA analysts be included as respondents for the deductive technique (Lopez, 1988). Contingent on this criterion, Lopez estimated the split half reliability for the TTA on an earlier version of the TTA, which included 33 traits. Across 100 jobs, analysts for each job were randomly allocated into two groups, then the correlation between the mean ratings given on each trait was calculated for the two groups. From this data set, the median split half reliability coefficient was 0.86 (Lopez, et al., 1981). On the basis of these findings, and under the constraints set by the organisation under scrutiny, a goal of nine participating analysts was set for the TTA analysis. Response rates were positive, with eight analysts returning questionnaires for the general sales and merchandiser positions, and six analysts returning questionnaires for the one-on-one sales position. These numbers were all within the limits set for achieving acceptable reliability in the TTA.

Respondents were selected on the basis of holding over six months experience in their positions, defining them as subject matter experts in their respective fields (Williams & Crafts, 1997). Biographical information pertaining to the TTA respondents was not obtained because of the likelihood that with the small numbers of respondents an individual could be personally identifiable on the basis of such information. Although specific information on the length of experience that the SME panel had was not obtained, the organisation under study only chose SMEs who had well over the six month criterion.

Summarising/Scoring Reponses to the TTA

Procedures for the scoring or summarising the results of the TTA are detailed in Lopez (1988). Certain criteria must be met for traits within the taxonomy to be deemed significant for inclusion in a selection plan. Firstly, only traits that are considered relevant by more than 40% of the analyst team are included in the final TTA analysis. Because other authors have suggested limiting the number of traits or dimensions that should be assessed in an AC to avoid cognitive overload (Arthur, et al., 2000; Lievens, 1998), it was decided that this criterion should be doubled. The number of traits selected was therefore reduced to an appropriate size by selecting those chosen as relevant by at least 80% of, rather than 40%, of the analyst team. This more stringent criterion also had possible added benefit of increasing the inter-rater agreement among the analyst group.

The next criterion used for deduction in the TTA is that pertaining to the levels of particular traits. It is assumed that any trait which occurs at level 0 for either standard or superior performance in a given role is at the same level of that trait that is "possessed by 90 percent of employable people" (Lopez, 1988, p. 888). Such traits are therefore not deemed vital for the particular position and are omitted from the selection plan. Lopez states that if there are major differences between analysts (e.g., one rater calls one particular trait a level 3 on superior performance whilst another analyst calls a level 0) reviewers should be reproached and questioned again to confirm the integrity of their judgements. There were 5 occasions in the present study where this was necessary, and generally, these inconsistencies were due to misunderstandings related to wording in the TTA questionnaire.

To identify which traits were most important, particularly at entry level into the respective positions, the final list of traits were subjected to the same procedure as the task list with a dichotomous (yes/no) screening item which probed whether the analyst panel was certain that the trait was relevant to the position, the criticality of the trait for proper performance on the job, and the importance of the possession of the trait upon entry to the job, all rated on generalised 6-point scales ranging from 0: "this trait was not relevant" to 5: "this trait was extremely important". A common theme that arose from the SME panel was the absence from the TTA of a trait concerning customer service orientation. Therefore, a suitable definition of customer service orientation was added for assessment in the new questionnaire. There was a unanimous agreement across the SME panel that a summarised version of the factors involved in customer service orientation provided by Saxe and Weitz (1982) reflected all of the elements that were endorsed by the organisation under scrutiny.

The time at which the second questionnaire was administered coincided with several major restructuring changes in the organisation under study. This may have contributed to a slightly lower return rate on the second trait analysis questionnaires. For the general sales and merchandiser positions, there were four respondents in total, and for the one-on-one sales position there were five respondents. The response rates were not deemed catastrophic, as they were still within the numbers of SMEs necessary to complete such analyses as suggested by Aamodt (1999) and Green and Stutzman (1986).

To compensate for the lower numbers of SMEs detailed above, a lower criterion for deduction of a trait was set at over 50% agreement for the relevance of a given trait. This was to allow for a larger number of traits to be subjected to scrutiny by the

management level SMEs mentioned later. As a result, most of the traits identified earlier were included for subsequent analysis, a result that was not surprising, as the set of traits, except for the trait relating to customer service orientation, had already been identified as being relevant by the TTA. The median score of the criticality and importance on entry of each trait was then calculated.

Presentation to the Managerial Level SME Panel

All task and trait level information was tabulated in order of importance as reported by the job incumbents. This information was then presented to a panel of managerial level SMEs across the different store locations. The managerial positions were general, in that all managers had experience managing the three positions under scrutiny. Fifteen managerial SMEs participated in this endeavour, and it was thought that such a presentation would provide a holistic view of the critical tasks and traits required to perform the job effectively at entry level. Additionally, the inclusion of managerial SMEs increased the number of respondents for each position, and therefore, theoretically, the ecological validity of the findings. Again, biographical information was not collected because of the confidentiality concerns outlined earlier. Biographical information was not thought to be vital to the replicability of the study. The essential criteria for selection as a managerial SME was at least six months experience as a manager for the three positions being assessed (Williams & Crafts, 1997, pp. 74-75).

The managerial SME panel were asked to read each task and each trait related to each of the three positions, and were asked to critique a) the relevance of the tasks/traits b) the omission of tasks/traits and c) the relative importance of the tasks/traits as rated by

the incumbents. On the basis of the suggestions of the SME panel, several modifications were made to the final task and competency information that would be used to guide the construction of the AC.

Classification and Extrapolation of Tasks into Dimensions

Tasks were classified into dimensions using the same process as was followed in the pilot for Study Three (see Appendix I). In a traditional dimension-specific AC, the tasks obtained from the task analysis are classified into dimensions. This involves identifying performance on the task with a dimension thought to underpin that performance (Ballantyne & Povah, 1995). An appropriate dimension was allocated to all tasks, in concurrence with SME suggestions in the present study (these dimensions are detailed in the following section). The current AC followed the growing body of literature suggesting that AC architects should limit the number of dimensions assessed (Gaugler & Thornton, 1989; Lievens & Klimoski, 2001; Sackett & Hackel, 1979) because some evidence suggests that using small numbers of performance dimensions may act to increase construct validity (Lievens, 1998). This issue possibly relates to cognitive overload as a factor that might interfere with the efficacy of rater judgements (Gaugler & Thornton, 1989; Reilly, et al., 1990). Upon reviewing the literature, Arthur et al. (2000) decided on a manageable set of nine performance dimensions, in conjunction with human information processing capacity. As previously mentioned, Gaugler and Thornton (1989) found evidence that the number of performance dimensions assessed should lie between five and seven. In the light of these findings, the number of

performance dimensions was limited to five per AC, thus aiming to minimise the cognitive load upon the assessors.

AC Task Ratings and Dimensions

Task checklists were developed for assessors to mark performance on the assessment exercises. These provided specific behavioural indicators of successful performance for each exercise. Assessors marked each task on a scale ranging from 1 (performance was certainly below standard) to 6 (performance was certainly above standard). Each task statement had a dimension name written next to it, to give the assessors guidance on which specific behaviours related to which dimension. Participants were rated on the following five dimensions:

Teamwork: The extent to which the individual works effectively and harmoniously with other team members.

Customer Focus: The extent to which the individual is concerned with customer needs, describes products accurately, matches presentations to the customer's interests, and attempts to assist customers to make satisfactory purchases.

Oral Expression: The extent to which the individual speaks grammatically and clearly in appropriate language and using appropriate gestures.

Tolerance: The extent to which the individual interacts effectively with people despite delicate, frustrating or tense situations that demand understanding, patience and empathy.

Comprehension: The extent to which the individual understands spoken and written, verbal, or behavioural language.

Each of these dimensions were assessed across all exercises. Although this is not frequently observed in ACs generally, the exercises were designed specifically to elicit behaviour pertaining to all dimensions so as to obtain as much data as possible pertaining to each dimension. Additionally, for ease of data analysis, such a design represented a fully crossed exercise by trait design. Dimensions were assessed on the same scale as behaviours. This standardization assisted comparisons in subsequent analyses.

AC Exercises

Three simulation exercises were employed to assess the tasks and dimensions in the AC. A brief description of these exercises follows:

Exercise 1, Approach Exercise: A group analysis exercise, in which candidates were presented with three situations where a hypothetical customer entered a store. The group's task was to plan the best method of approach that should be applied to the each customer.

Exercise 2, Closing Exercise: A strategic group discussion exercise, where candidates were presented with six different written scenarios for each of which they had to choose appropriate ways of closing a sale.

Exercise 3, Returns Exercise: A group analysis exercise where candidates were presented with four situations where a hypothetical customer arrived to return goods to a store. The group had to come to a consensus as to the best method of handling the returns issues presented by individual customers.

Evaluation Approach

Several studies have sought to assess the relative efficacy of evaluating performance dimensions after the completion of each exercise, or waiting until the completion of the entire AC before making an evaluation of the dimensions concerned (within-exercise rating versus within-dimension rating). As previously discussed, the evidence for the efficacy of one approach over the other remains unclear (Harris, et al., 1993; Silverman, et al., 1986). As the two approaches appear to contribute relatively little to the facilitation of the construct validation of the AC process, it was decided that the within-exercise approach would be used. As previously argued, if the dimensions in ACs are conceptualised as being relatively stable, then theoretically, they should stand up to being rated in individual exercises without the necessity of having to wait until after the process is finished (refer Campbell & Fiske, 1959). Also, leaving an assessment such as this until the end of the process may act to increase the level of error associated with the behavioural judgements that are thought to be manifestations of the underlying trait, because at the end of the AC, assessors may forget what behaviour they saw whilst observing a given exercise.

Assessor Training and the Assessment Procedure

Assessors were trained on the AC exercises using a mixture of behavioural observation training (Ballantyne & Povah, 1995) coupled with guidance on how to use behavioural checklists to assist the process (Lowry, 1997). It has been suggested that frame of reference training (Bernardin & Buckley, 1981; Sulsky & Day, 1992) is efficacious for the appraisal of human performance (Murphy & Cleveland, 1995) as well

as tending to increase the construct and criterion validity of AC ratings (Arthur, et al. 2000; Lievens, 1998; Schleicher, 1999). The present study used frame of reference training as an integral aspect of the procedure.

In sequence, the process firstly involved a general explanation of the AC process, and the benefits to the organisation of utilising this procedure. A general description of the two types of judgement that would be required was given, being the observation of behaviours, and arising from those behaviours, the inference of underlying traits. In congruence with the guidelines of Ballantyne and Povah (1995), assessors were then shown how to assess behaviours with no construct inference. This process involved observation and recording a score on a behavioural checklist (Lowry, 1997). This procedure increased objectivity, and guided the scoring process, allowing for numerical rating data to be obtained from the behaviour that was observed. The inference of ability traits was made over and above these behavioural ratings at the next stage of classification. Here, assessors were shown examples of the behaviours theoretically constituting evidence for the presence of each hypothetical trait.

Each previously rated behaviour was thus denoted as being a possible underlying indicator for a superordinate dimension. Assessors gave an inferred score for these dimensions on a 5-point scale (Ballantyne & Povah, 1995). Assessors were then trained in the consensus discussion procedure for dimensional ratings, where at the conclusion of the AC, all assessors presented evidence and critically discussed the ratings they had obtained to form OARs for each participant. Assessors considered the participants individually, and assessed their performance on each dimension individually.

Assessors were trained on assessing participants using a frame of reference training procedure (Lievens, 1998; Lievens, 2001a). This involved a training session with assessors that first covered some basic principles in assessing behaviour, and familiarised the assessors with the exercises and the rating instruments that would be used. The FOR component of the training initially involved presenting the assessors with the definitions of the dimensions. Then, incorporating the information from the behavioural checklists, the assessors engaged in discussion concerning the behaviours that were associated with different levels of dimensions.

The group was then presented with a written exercise listing 20 behavioural incidents that reported people behaving as they might in an AC. These incidents were devised in accordance with the suggestions of the SME group. The assessor's task was to assign each incident to one of the performance dimensions, and to assign a performance rating to each of the incidents. The ratings that had been allocated were then discussed as a group, in relation to the responses to the classification of the behavioural incidents previously given by the SME group. Raters were encouraged to give justifications for their ratings. Raters were then given experience on rating the actual AC, used in this study, with role-playing participants. All raters were instructed to rate the same participants, so that their ratings could be compared later. After each exercise had been completed, the ratings awarded were displayed using a projector linked to a computer. Discussions focused on the scores that had been awarded for each behavioural item on the behavioural checklist, and each dimensional rating. In particular, the focus was on scores that were notably deviant from others. This process was conducted with reference to the mean and standard deviations of the ratings for each assessor on each dimension,

which were displayed graphically. This was done in order to promote the development of a shared schema with respect to good versus poor performance on a given exercise.

Procedure

The centre was run according to a schedule where each group of participants performed each exercise in turn, whilst the allocation of participants to assessors was assigned systematically on each exercise. This was performed in such a way that each participant was assessed by a different assessor in each exercise, as suggested by Lievens (1998). At the conclusion of this process, the assessors calculated average ratings to derive an OAR, as in Pynes and Bernardin (1992) and as approved by the International Task Force on Assessment Center Guidelines (2000). Indeed, because large numbers of individuals were being assessed, it was not practical to engage in integration discussions about each participant. Note, however, that integration discussions are used in ACs to determine OARs. OARs were not included in the analyses employed in this study, as the research question concerned the allocation of ratings, as opposed to the derivation of OARs. Thus, the final mechanical integration of ratings in this AC should not have affected the observed ratings included for analysis.

Results

While 240 individuals participated in the AC, the company under scrutiny was able to provide ratings for only 199. The location of the 41 missing results could not be ascertained, although the company stated that they were lost on a random basis, and did not relate to any particular subset of individuals. Of the 199 results that were received, 187 were deemed usable. Non-usable candidate rating sheets were those in which the ratings of entire exercises were missing for various reasons, such as data sheets being mislaid or candidates leaving the AC. The data in these sheets were considered too incomplete for inclusion in the analysis. Algorithm

The data from the remaining 187 participants for Study Two were imputed for missing values using EM (expectation maximisation), which uses an iterative process, by which to estimate missing values. This method was recommended by Gold and Bentler (2000) for optimal data substitution, regardless of sample size, proportion of missing data, and distributional characteristics. Two sets of data were of interest; a set of task-specific data and a set of dimension-specific data. For the task-specific data, there were a total of seven missing values out of a possible total of 7480 ratings. This constituted a 99% response rate for the task-specific rating scales. For the dimension-specific data, there were no missing values out of a total of 2805 ratings. The high response rates here were attributable, perhaps, to reiteration in training that assessors should provide a response to every item listed on the ratings scales.

Table 32 shows the exercise grand means and standard deviations for the task-specific AC. Under the task-specific approach, performance on particular exercises is considered the most important unit of measurement. All mean scores vacillated around the 4th point on the rating scale. The grand mean rating for the last exercise

was slightly lower than the others at 3.65. Standard deviations for the task-specific ratings were also fairly comparable. Table 33 shows the grand means and standard deviations for the dimension-specific AC. Under the dimension-specific approach, performance on particular dimensions is considered the most important unit of

Table 32

Grand Means and SDs of the Behavioural Ratings (Within Exercises) in the Task-Specific AC

Exercise	*M*	*SD*
Approach	4.02	1.35
Closing	4.14	1.29
Returns	3.65	1.41

Table 33

Grand Means and SDs of the Dimension Ratings (Across Exercises) in the Dimensi Specific AC

Dimension	*M*	*SD*
Comprehension	4.13	1.22
Oral Expression	4.00	1.36
Tolerance	4.10	1.28
Teamwork	3.70	1.35
Customer Focus	3.73	1.36

measurement. Like the task-specific AC, average dimensional ratings centred around the 4th point on the rating scale. Standard deviations for the dimension-specific ratings were fairly comparable, at less than 1.50.

Generalizability Study

The present study employed Generalizability Theory (G theory) (Brennan, 2001a; Cronbach, Gleser, Nanda & Rajaratnam, 1972) to analyse data. The reader who is unfamiliar with G theory is directed to Appendix II. Although statistical significance is not generally considered to be of importance in G theory (Brennan, 2000), the confidence *limits* within which one computes estimates of components of variance can be calculated using confidence intervals designed specifically for variance component estimates (Brennan, 2001a). Such confidence intervals cannot be theoretically justified for designs that are unbalanced with respect to nesting (Brennan, 2001b). The task-specific design in Study Three was slightly unbalanced in this regard, in that exercise one contained 14 behavioural checklist items, while exercises two and three contained 13 items each. Thus, in the interests of gaining information on the confidence associated with the estimated variance components, one item was removed from exercise one to create a balanced design. It was decided that item three of exercise one should be omitted, because it shared an identical overall average value, and not vastly different SDs, across all 187 subjects with item four within the same exercise (mean for item three = 3.36, SD = 1.46, mean for item four = 3.36, SD = 1.37). Additionally, items three and four correlated notably ($r_{3,4}$ = .80, $p < .001$) Thus, it was reasoned that with another item within the same exercise that shared similar central tendency characteristics, item three would not contribute a great deal of variance to scores within exercise one. A Generalizability study (G

study) was performed for all balanced designs in the present study, plus an extra G study for the complete, unbalanced task-specific design (i.e., with item three of exercise one included), without the calculation of confidence intervals.

G studies utilise variance components models that are derived from the mean squares calculated in factorial ANOVAs. The goal in a G study is to identify important facets that may contribute to variance in scores (Shavelson & Webb, 1991). The selection of facets is, however, restricted to the constraints associated with ANOVA analyses. ANOVA models are less restrictive under the highly controlled circumstances in which levels of all facets are associated with all levels of other facets (i.e., a fully crossed design). In situations where levels of a particular facet are not always systematically allocated in a design, it becomes impossible to disentangle a specific source of variation for that facet. In the case of the present study, the effect for assessors was not always systematically allocated in the design. As the present study was a field study, the gain of greater ecological validity comes with the cost of lessened control over the variables of study.

The effects of differences between raters both within and between the ACs were not thought to be of great concern to the primary aims in the present study, because the same raters were used for the same participants across the two ACs and the principal purpose of the study was to compare alternative types of AC. Thus, the assignment of raters was such that the effect attributable to raters would be held constant over the task-specific and dimension-specific administrations. Assessors were allocated on a rotational basis, as suggested by Lievens (1998), as an attempt to randomise rater error. However, during the course of the AC, assessor allocation was not always systematic because the current field study did not have strict control over all the variables under scrutiny. The complexities of the unsystematic nesting of

assessors disallowed their inclusion in the present G study. To gain an estimate of interrater reliability, Equation 1,1 from Shrout and Fleiss (1979; see Tables 35 and 36 for Equation 1,1) was employed for each AC. Equation 1,1 was relevant to the present sample because each participant was rated by a random combination of assessors who were selected from a larger population of judges.

Specific facets were included in the G study that were instrumental in addressing the research issue at hand. The task-specific AC constituted a partially nested design, in that each exercise had its own specific set of items. The facets included in the task-specific process included exercises (x), and items nested within exercises ($i{:}x$). Variability attributable to the object of measurement, persons (p) was also estimated. All interaction terms were analysed. The dimension-specific AC employed a fully crossed design incorporating the facets exercises (x) and dimensions (d). An estimate of the variability attributable to the object of measurement, persons (p) was also estimated. All interaction terms were analysed. The above describes standard practice in G studies of this nature (Shavelson & Webb, 1991).

Table 34 shows the G study for the comparison between the balanced task-specific and the dimension-specific ACs. All variance components and confidence intervals were computed using urGenova, Version 2.1 (Brennan, 2001b). Listed for each type of AC are the object of measurement, facets, and interactions (effects), degrees of freedom (*df*), variance component estimates (VC), 90% confidence intervals and the percent of explained variance (explained variance %) as a heuristic for identifying the proportional contribution of various facets to variation in scores (Shavelson & Webb, 1991).

In the task-specific approach, the px interaction was a comparatively high contributor, explaining 30.1% of the variance in the model. A proportionately high px

Table 34

Generalizability Study Comparing a Task-Specific with a Dimension-Specific AC in a Repeated Measures Design for the Organisational Sample

Task-Specific AC					Dimension-Specific AC				
Effect	*df*	VC	90% Confidence Intervals	Explained Variance (%)	Effect	*df*	VC	90% Confidence Intervals	Explained Variance (%)
p(persons)	186	0.5174	0.4049 < VC < 0.6618	27.4	*p*(persons)	186	0.5847	0.4560 < VC < 0.7495	31.9
x(exercises)	2	0.0579	0.0100 < VC < 1.3461	3.1	*x*(exercises)	2	0.0580	0.0164 < VC < 1.2113	3.2
i(items):*x*	36	0.1097	0.0765 < VC < 0.1716	5.8	*d*(dimensions)	4	0.0395	0.0154 < VC < 0.2314	2.2
px	372	0.5789	0.5099 < VC < 0.6623	30.7	*px*	372	0.6143	0.5355 < VC < 0.7095	33.6
pi:*x*,*e*	6696	0.6234	0.6061 < VC < 0.6415	33.0	*pd*	744	0.0328	0.0138 < VC < 0.0533	1.8
					xd	8	0.0030	0.0003 < VC < 0.0140	0.2
					pxd,*e*	1488	0.4980	0.4693 < VC < 0.5295	27.2

Note: Confidence intervals were calculated using the Ting et al. (1990) procedure described in Brennan (2001a). Ting et al's procedure is recommended for random, balanced designs so as to avoid the computation of inaccurately wide intervals.

interaction in the task-specific approach is defined by variation in the candidate's performance according to different situations (exercises) presented to them. In the dimension-specific approach, the *px* interaction was also comparatively high at 33.5%. Again, this interaction reflects the extent to which candidate performance varied across exercises.

The interaction term *pd*, in the dimension-specific AC, reflects the extent to which dimensions are useful for discriminating between persons (Lievens, 2001a; 2001b). This interaction term explained 1.8% of the variance in the dimension specific model. Additionally, the effect for the object of measurement, *p*, was estimated for the task-specific approach, and explained 27.4% of the total variance. The object of measurement, *p*, for the dimension specific approach was marginally higher, and explained 31.9% of the variance in scores. The terms *pi:x,e* and *pxd,e* in the task-specific and dimension-specific processes, respectively, are difficult to interpret purely as they contain the interactions between all facets and the object of measurement in the model, plus undifferentiated random error.

Confidence intervals are presented in Table 34, and are also graphically represented in Figure 2 for the task-specific model, and Figure 3 for the dimension-specific model. All confidence intervals were calculated using the method suggested by Ting, Burdick, Graybill, Jeyaratnam, and, Lu (1990), which is generally recommended for random, balanced designs so as to avoid the computation of inaccurately wide intervals for variance component estimates (Brennan, 2001a).

G theory acknowledges that in practice, relative and absolute decisions are often made about individuals on the basis of a psychological measure. A relative decision is one in which the performance of individuals are compared with other individuals (e.g., norm comparisons present relative decisions where people are

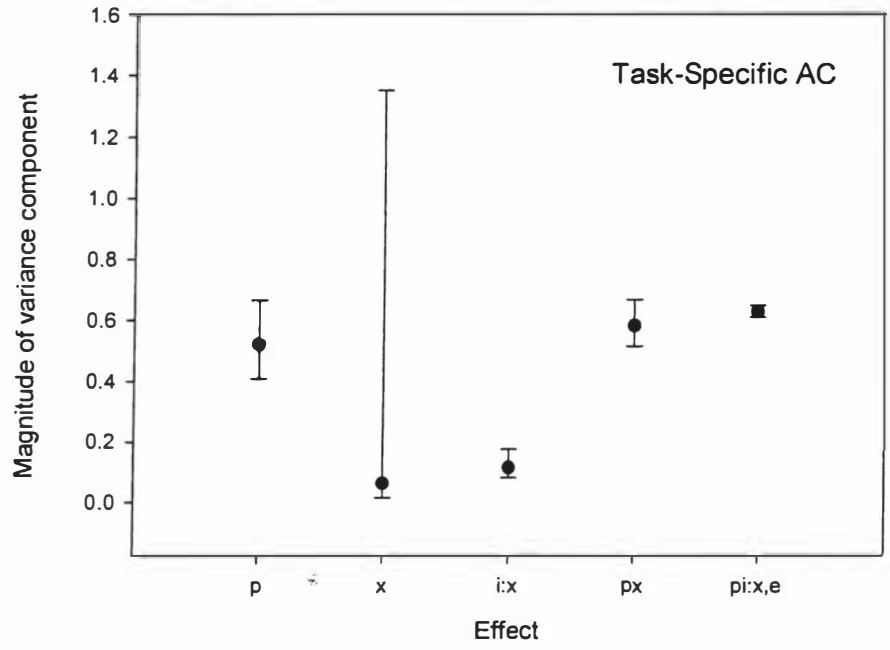


Figure 2. Variance Components and Confidence Intervals for Each Effect and Interaction in the Task-Specific AC.

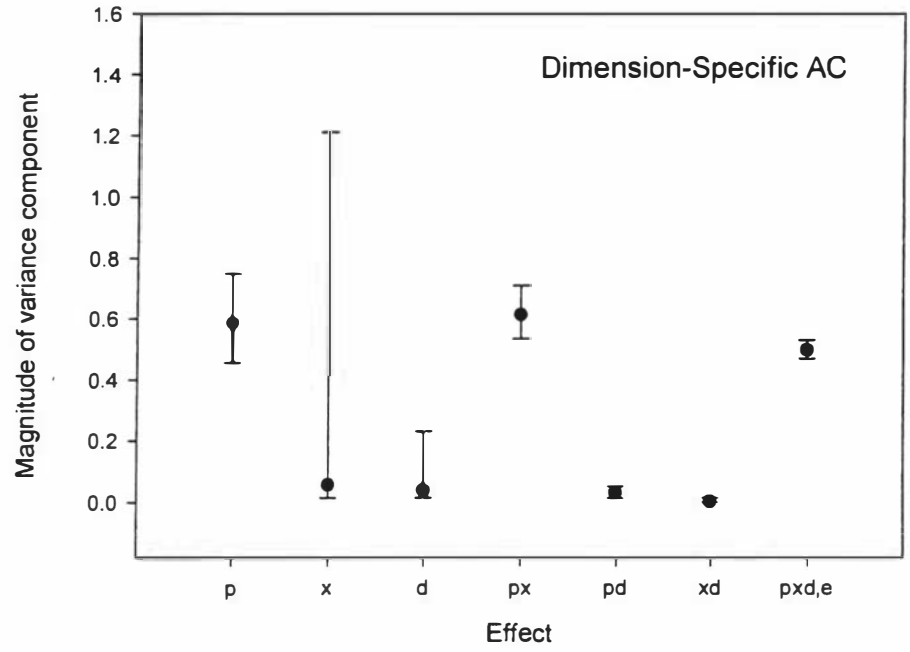


Figure 3. Variance Components and Confidence Intervals for Each Effect and Interaction in the Dimension-Specific AC.

compared with one another). An absolute decision is one in which a certain cut-off criterion is employed (e.g., a pass or fail criterion for employment decisions). G theory provides two coefficients for the purposes of relative and absolute decisions that are analogous to reliability coefficients in classical test theory. Tables 35 and 36 provide the equations and calculations, for both types of AC, of σ^2_{Rel} (relative error; all of the effects in the G study that contribute variance to relative decisions), σ^2_{Abs} (absolute error; all of the effects in the G study that contribute variance to absolute

Table 35

Relative and Absolute Error, Generalizability and Phi Coefficients and Interrater Reliability for the Balanced Task-Specific AC

Index	Result
$\sigma^2_{Rel} = \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_{pi:x,e}}{n_{i:x}n_x}$	0.20
$\sigma^2_{Abs} = \frac{\sigma^2_{x}}{n_x} + \frac{\sigma^2_{i:x}}{n_{i:x}} + \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_{pi:x,e}}{n_{i:x}n_x}$	0.22
$E\rho^2_{Rel} = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{Rel})}$	0.72
$\phi = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{Abs})}$	0.70
$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS}$	0.93

Table 36

Relative and Absolute Error, Generalizability and Phi Coefficients and Interrater Reliability for the Dimension-Specific AC

Index	Result
$\sigma^2_{\text{Rel}} = \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_d}{n_d} + \frac{\sigma^2_{pxd,e}}{n_x n_d}$	0.24
$\sigma^2_{\text{Abs}} = \frac{\sigma^2_x}{n_x} + \frac{\sigma^2_d}{n_d} + \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_{pd}}{n_d} + \frac{\sigma^2_{xd}}{n_x n_d} + \frac{\sigma^2_{pxd,e}}{n_x n_d}$	0.27
$E\rho^2_{\text{Rel}} = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{\text{Rel}})}$	0.71
$\phi = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{\text{Abs}})}$	0.68
$\text{ICC}(1,1) = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (k-1)\text{WMS}}$	0.82

decisions), $E\rho^2_{\text{Rel}}$ (the Generalizability or G coefficient; for relative decisions), and ϕ (Phi coefficient; for absolute decisions). Tables 35 and 36 also provide equation ICC 1,1 from Shrout and Fleiss (1979) as an estimate of interrater reliability across the two types of AC.

Table 37 presents the variance component estimates for the full, unbalanced task-specific approach with item three of exercise one included in the analysis. As expected, the pattern of findings were similar to those in Table 34 for the task-specific approach. Thus, it appears unlikely that item three of exercise one contributed much

Table 37

Generalizability Study Showing the Results of the Unbalanced Task-Specific AC for the Organisational Sample

Effect	*df*	VC	90% Confidence Intervals	Explained Variance (%)
p(persons)	186	0.51073	*	26.8
x(exercises)	2	0.05125		2.7
i(items):*x*	37	0.11758		6.2
px	372	0.58488		30.7
pi:x,e	6882	0.63804		33.5

* Confidence intervals were not provided for the task-specific procedure in this case, because the specification of a confidence interval for an unbalanced design is inappropriate (Brennan, 2001b).

variance to the scores, and could safely be regarded as a redundant source of variation.

Factor Analysis

In the tradition of several other studies on AC ratings, including the seminal paper by Sackett and Dreher (1982), a factor analysis was employed to evaluate the measurement models presented in the task-specific and the dimension-specific ACs (i.e., to provide what might be viewed as a more traditional perspective on the same data). SPSS ver. 11 was employed to produce communalities and factor loadings for both types of AC. The same raw data as in the G study was used as input for the

factor analysis. Note that the full unbalanced data set was used for the task-specific AC. Principle axis factoring was employed as the method of extraction. In principle axis factoring, communality estimates are derived through an iterative procedure, using squared multiple correlations of each variable with all other variables as the starting point. The goal of principle axis factoring is to extract maximum orthogonal variance from the data with the extraction of each successive factor (Tabachnick & Fidell, 1983). Principle axis factoring is widely employed, and was also used in Sackett and Dreher's original study. The present study used .40 as a criterion for an admissible factor loading, in congruence with Comrey and Lee (1992), who suggest that factor loadings of .45 upwards are fair indicators of the overlap between a variable and a factor. All factor loadings are displayed in the analyses, however, as different researchers set different criteria for acceptable factor loadings (Tabachnick & Fidell, 1983). Varimax rotation was employed to encourage simple structure in the ratings, to assist for comparison purposes with seminal pieces on this topic (e.g., Sackett and Dreher, 1982) and for ease of interpretation. As practitioners ideally strive for simple factor structure as the basis for AC ratings, varimax was seen as most appropriate. Direct oblimin was also employed as a rotational method across both the task-specific and the dimension-specific models to allow for correlation among factors, and for comparison purposes.

Varimax Rotation

Three factors were extracted for the task-specific AC. This is because under the behavioural task-specific paradigm, each of the three exercises in the AC was viewed as a stand-alone work sample of behaviour. Table 38 shows the results for the

Table 38

Rotated Factor Matrix for the Task-Specific AC Ratings

		Factor Loadings					
Exercise	Item	1	2	3	Communality	*M*	*SD*
Approach	1	.24	.01	**.80**	.72	4.29	1.17
Approach	2	.20	.18	**.71**	.58	4.14	1.35
Approach	3	.09	.13	**.70**	.52	3.36	1.46
Approach	4	.13	.17	**.76**	.63	3.36	1.37
Approach	5	.19	.14	**.71**	.57	4.28	1.32
Approach	6	.12	.20	**.58**	.39	4.18	1.12
Approach	7	.18	.10	**.68**	.50	4.27	1.24
Approach	8	.24	.10	**.79**	.69	3.71	1.39
Approach	9	.14	.15	**.78**	.65	4.22	1.26
Approach	10	.14	.12	**.74**	.57	3.58	1.44
Approach	11	.32	.13	**.74**	.67	4.07	1.34
Approach	12	.29	.14	**.77**	.69	3.89	1.31
Approach	13	.12	.11	**.71**	.53	4.43	1.19
Approach	14	.09	.05	**.64**	.42	4.56	1.25
Closing	15	.28	**.75**	.21	.69	3.89	1.30
Closing	16	.31	**.78**	.25	.76	3.91	1.37
Closing	17	.18	**.79**	.26	.72	4.01	1.38
Closing	18	.33	**.75**	.19	.70	3.78	1.35
Closing	19	.16	**.79**	.17	.68	4.21	1.29
Closing	20	.24	**.80**	.17	.72	3.72	1.36
Closing	21	.22	**.75**	.13	.63	4.20	1.31
Closing	22	.22	**.78**	.19	.68	3.90	1.32
Closing	23	.18	**.71**	.10	.54	4.39	1.27
Closing	24	.10	**.78**	.06	.62	4.38	1.16
Closing	25	.17	**.68**	.03	.49	4.33	1.08
Closing	26	.07	**.71**	.08	.52	4.67	1.09
Closing	27	.10	**.83**	.08	.70	4.39	1.17
Returns	28	**.85**	.23	.17	.80	3.50	1.35
Returns	29	**.87**	.16	.11	.80	3.35	1.34
Returns	30	**.81**	.15	.13	.69	3.51	1.40
Returns	31	**.82**	.22	.22	.77	3.75	1.35
Returns	32	**.79**	.20	.18	.70	3.65	1.40
Returns	33	**.84**	.18	.23	.80	3.42	1.36
Returns	34	**.79**	.19	.23	.72	3.02	1.39
Returns	35	**.72**	.26	.20	.63	3.90	1.33
Returns	36	**.77**	.25	.18	.69	3.35	1.48
Returns	37	**.79**	.23	.24	.74	3.82	1.43
Returns	38	**.78**	.24	.25	.73	3.62	1.40
Returns	39	**.66**	.19	.23	.52	4.15	1.28
Returns	40	**.67**	.12	.24	.53	4.43	1.30
Eigenvalue		9.09	8.33	8.26			
% of variance explained		22.73	20.84	20.64			

task-specific AC. In congruence with the results of the G study, clear loadings on exercises were found for the task-specific AC. Table 39 shows the results for the dimension-specific AC. Five factors were extracted for the dimension-specific AC, because five dimensions were included in the assessment. Relatively clear factor loadings on exercises, that is, three exercise factors were evident, in congruence with the G study perspective on these same data.

Table 39

Rotated Factor Matrix for the Dimension-Specific AC Ratings

		Factor Loadings							
Dimension	Exercise	1	2	3	4	5	Communality	*M*	*SD*
Comprehension	Approach	.28	.15	**.63**	**.58**	.06	.83	4.20	1.15
Oral Expression	Approach	.32	.17	**.72**	.27	-.03	.73	3.95	1.32
Tolerance	Approach	.10	.18	**.80**	.11	.10	.70	4.24	1.21
Teamwork	Approach	.19	.21	**.90**	-.17	.00	.91	3.80	1.30
Customer Focus	Approach	.20	.20	**.77**	.03	-.07	.66	3.91	1.35
Comprehension	Closing	.14	**.75**	.10	.18	.39	.78	4.38	1.13
Oral Expression	Closing	.21	**.78**	.17	.01	.05	.70	4.07	1.30
Tolerance	Closing	.17	**.79**	.15	.01	.01	.68	4.18	1.25
Teamwork	Closing	.29	**.86**	.18	.04	-.20	.89	3.88	1.34
Customer Focus	Closing	.23	**.81**	.22	.00	-.05	.75	3.89	1.32
Comprehension	Returns	**.77**	.20	.20	.10	-.11	.70	3.80	1.31
Oral Expression	Returns	**.81**	.24	.20	.07	-.08	.76	3.65	1.43
Tolerance	Returns	**.78**	.17	.18	.02	.20	.71	3.88	1.36
Teamwork	Returns	**.89**	.20	.21	.03	.06	.87	3.41	1.38
Customer Focus	Returns	**.84**	.22	.15	.03	-.02	.78	3.40	1.37
Eigenvalue		3.85	3.56	3.27	.50	.28			
% of variance explained		25.69	23.76	21.81	3.36	1.86			

Direct Oblimin Rotation

In order to determine the appropriate delta parameter for direct oblimin rotation, PsWin ver. 2.0.1 was utilised (Barrett, 1996). PsWin has the capability to assess the delta parameter that will maximise simple structure in a direct oblimin rotation. In both the task-specific and the dimension-specific data, a delta value of zero was assessed as optimal. For the task-specific AC, three factors were extracted. For the dimension specific AC, five factors were extracted. Table 40 shows the direct oblimin results for the task-specific AC. Table 41 shows the direct oblimin results for the dimension-specific AC. Five factors were extracted for the dimension-specific AC as five dimensions were included in the assessment. Relatively clear factor loadings on exercises were obtained for the task-specific approach, and although less clear than the varimax rotation, factor loadings were tending to load onto exercises in the dimension-specific approach also. The exceptions to clear exercise effects are summarized in the following. Oral expression tended to bleed across factors four and five in the approach exercise, and comprehension tended to remain in factor four for this exercise. Comprehension bled across factors one and two in the closing exercise.

Table 40

Rotated Pattern Matrix for the Task-Specific AC Ratings

		Factor Loadings					
Exercise	Item	1	2	3	Communality	*M*	*SD*
Approach	1	**.82**	.08	-.10	.72	4.29	1.17
Approach	2	**.72**	-.06	-.04	.58	4.14	1.39
Approach	3	**.75**	-.03	.08	.52	3.36	1.46
Approach	4	**.80**	-.06	.05	.63	3.36	1.37
Approach	5	**.73**	-.01	-.03	.57	4.28	1.32
Approach	6	**.59**	-.12	.04	.39	4.18	1.12
Approach	7	**.70**	-.01	-.03	.50	4.27	1.24
Approach	8	**.81**	-.05	-.08	.69	3.71	1.39
Approach	9	**.81**	-.03	.04	.65	4.22	1.26
Approach	10	**.77**	-.00	.02	.57	3.58	1.44
Approach	11	**.73**	.02	-.17	.67	4.07	1.34
Approach	12	**.77**	.02	-.13	.69	3.89	1.31
Approach	13	**.75**	-.00	.04	.53	4.43	1.19
Approach	14	**.69**	.05	.05	.42	4.56	1.25
Closing	15	.07	**-.74**	-.12	.69	3.89	1.30
Closing	16	.10	**-.76**	-.13	.76	3.91	1.37
Closing	17	.14	**-.80**	.03	.72	4.01	1.38
Closing	18	.03	**-.73**	-.18	.70	3.78	1.35
Closing	19	.05	**-.82**	.03	.68	4.21	1.29
Closing	20	.03	**-.81**	-.06	.72	3.72	1.36
Closing	21	-.00	**-.77**	-.06	.63	4.20	1.31
Closing	22	.10	**-.79**	-.04	.68	3.90	1.32
Closing	23	-.03	**-.74**	-.02	.54	4.39	1.27
Closing	24	-.05	**-.83**	.07	.62	4.38	1.16
Closing	25	-.09	**-.71**	-.04	.49	4.33	1.08
Closing	26	-.02	**-.76**	.09	.52	4.67	1.09
Closing	27	-.04	**-.88**	.08	.70	4.39	1.17
Returns	28	-.05	-.03	**-.90**	.80	3.50	1.35
Returns	29	-.11	.03	**-.95**	.80	3.35	1.34
Returns	30	-.07	.03	**-.87**	.69	3.51	1.40
Returns	31	.02	-.02	**-.85**	.77	3.75	1.35
Returns	32	-.01	-.02	**-.83**	.70	3.65	1.40
Returns	33	.03	.02	**-.89**	.80	3.42	1.36
Returns	34	.04	-.00	**-.83**	.72	3.02	1.39
Returns	35	.02	-.10	**-.73**	.63	3.90	1.33
Returns	36	-.01	-.08	**-.80**	.69	3.35	1.48
Returns	37	.05	-.04	**-.82**	.74	3.82	1.43
Returns	38	.06	-.05	**-.80**	.73	3.62	1.40
Returns	39	.07	-.03	**-.67**	.52	4.15	1.28
Returns	40	.10	.07	**-.71**	.53	4.43	1.30
Eigenvalue		11.38	11.44	13.02			
% of variance explained		41.34	12.50	10.35			

Table 41

Rotated Pattern Matrix for the Dimension-Specific AC Ratings

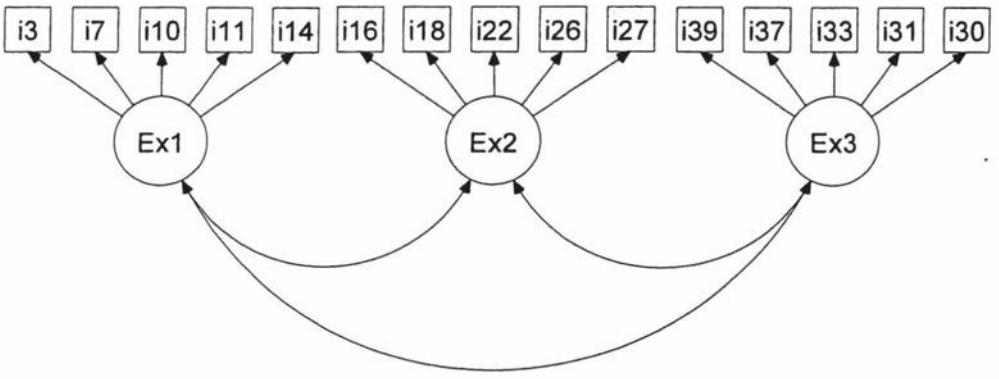
		Factor Loadings							
Dimension	Exercise	1	2	3	4	5	Communality	*M*	*SD*
Comprehension	Approach	-.02	-.01	-.08	**.80**	.11	.83	4.20	1.15
Oral Expression	Approach	.05	-.02	-.14	**.42**	**.44**	.73	3.95	1.32
Tolerance	Approach	-.09	-.03	.07	.21	**.70**	.70	4.24	1.21
Teamwork	Approach	-.00	-.05	-.04	-.13	**.99**	.91	3.80	1.30
Customer Focus	Approach	.08	-.03	-.04	.12	**.70**	.66	3.91	1.35
Comprehension	Closing	**-.40**	**-.71**	-.01	.15	-.10	.78	4.38	1.13
Oral Expression	Closing	-.06	**-.80**	-.04	-.04	.05	.70	4.07	1.30
Tolerance	Closing	-.02	**-.83**	.02	-.02	.03	.68	4.18	1.25
Teamwork	Closing	.21	**-.92**	-.05	.04	-.02	.89	3.88	1.34
Customer Focus	Closing	.05	**-.83**	-.03	-.03	.09	.75	3.89	1.32
Comprehension	Returns	.14	-.05	**-.77**	.11	-.04	.70	3.80	1.31
Oral Expression	Returns	.11	-.08	**-.81**	.05	-.02	.76	3.65	1.43
Tolerance	Returns	-.19	.06	**-.86**	-.05	.06	.71	3.88	1.36
Teamwork	Returns	-.04	-.02	**-.94**	-.01	.04	.87	3.41	1.38
Customer Focus	Returns	.04	-.04	**-.87**	-.01	-.02	.78	3.40	1.37
Eigenvalue		.40	5.04	5.46	3.47	4.34			
% of variance explained		47.46	12.82	11.87	2.77	1.57			

Confirmatory Factor Analysis

AMOS (version 4) was employed to evaluate the fit of three models that reflected the varying designs of AC mentioned or implied in this study. Each model is first represented graphically, then the associated factor loadings, together with goodness-of-fit indices, are presented in tabulated form. Raw data from the AC ratings was used as input for AMOS throughout the confirmatory factor analyses (CFA).

Model One: The Abridged Task Specific Model

The task-specific model was tested first (see Figure 4). The task-specific model tested was an abridged version of the observed model summarized in the previous analyses. Items were removed from the task-specific CFA model in light of cautions surrounding sample size (relative to the number of parameters estimated) when employing structural equation models (SEM) (Bollen, 1989; Klem, 2000). As a rule of thumb, Bentler and Chou (1987) recommended that at a minimum, five cases should be present per parameter estimated, and it is recommended that there should be 10 cases per parameter. The full task-specific model (i.e., without items removed) contained 83 parameters, and therefore would have required a sample size of 415 at a bare minimum. Thus, the sample size of 187 in the present study fell well short of



Note: Ex1 = approach exercise; Ex2 = closing exercise; Ex3 = returns exercise. The observed variables 'i3' through to 'i39' represent the randomly selected behavioural items associated with each exercise.

Figure 4. Model One: Abridged Task-Specific CFA Model.

this criterion. It was decided, therefore, that for an abridged version of the task-specific model, five task-specific items should be retained for each exercise to attempt a reasonable comparison with the dimension-specific model. This number was considered suitable because it allowed direct comparison with the dimension-specific AC, which employed five trait judgements per exercise. Items were retained in the full task-specific AC on a random basis to avoid bias in the selection of items to remove or retain. The random number table in Coolican (1999, p. 448) was employed for this purpose. Figure 4 shows the items that were retained in the abridged model. The resulting model therefore comprised three latent, and fifteen observed variables. The three latent variables in this case represented behavioural performance on exercises. Table 42 presents the standardised parameter estimates for the model represented in Figure 4. It was found that the mean values from the abridged data set and the mean values from the total data set were strongly correlated ($r = .99, p < .001$).

The model parameters shown in Table 42 indicate relatively clear factor loadings on exercises for the task specific model. This result was consistent with those reported in the generalizability study, and the factor analysis. Table 43 shows selected goodness-of-fit indices for the task-specific model. The ratio of case numbers to the number of parameters in the model was within the minimum limits suggested by Bentler and Chou (1987). Model One contained 33 parameters and 187 cases. Note that the samples employed in this entire study were still small when considering Bentler and Chou's suggestions. The reader is therefore cautioned that certain goodness-of-fit indices tend to underestimate fit when the sample size is small (Byrne, 2001, MacCallum, Browne & Sugawara, 1996).

Table 42

Standardised Factor Loadings for Model One: The Abridged Task-Specific CFA Model

	Exercises		
Behavioural Items	Ex1	Ex2	Ex3
i3	.61		
i7	.63		
i10	.77		
i11	.89		
i14	.64		
i16		.92	
i18		.89	
i22		.76	
i26		.58	
i27		.75	
i30			.70
i31			.86
i33			.88
i37			.86
i39			.80

Note: Ex1 = approach exercise; Ex2 = closing exercise; Ex3 = returns exercise. The observed variables 'i3' through to 'i39' represent the randomly selected behavioural items associated with each exercise.

Overall, goodness-of-fit indices presented in Table 43 were suggestive of a reasonable fit for the abridged task-specific model. GFI and AGFI indices were reasonable in the present study (Byrne, 2001). The CFI and TLI indices should approach .95 (Byrne, 2001) as a rule of thumb, and in this study, the CFI and TLI were again reasonable. Browne and Cudeck (1993) suggested that RMSEA values as high as .08 indicate a reasonable fit, thus the RMSEA point estimate of .079 was also suggestive of a reasonable fit.

Table 43

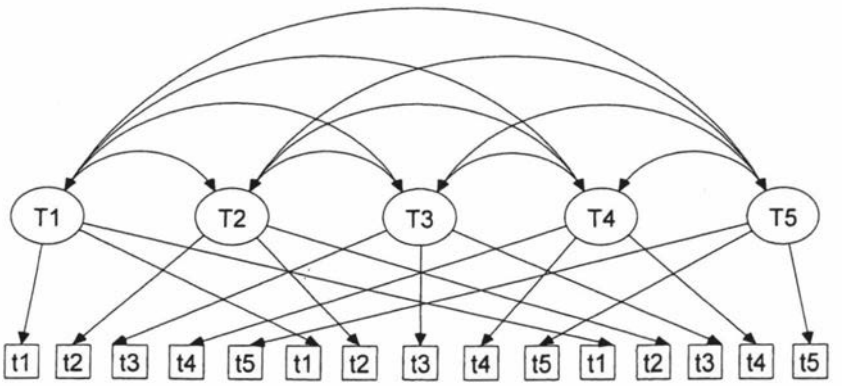
Selected Goodness-Of-Fit Indices for Model One: The Abridged Task-Specific CFA Model

Index	Point Estimate
GFI	.880
AGFI	.835
CFI	.940
TLI	.928
RMSEA*	.079

*90% Confidence Interval for RMSEA (.064 < RMSEA < .095)

Model Two: The Dimension-Specific CFA Model

The second model tested was the dimension specific model (see Figure 5). Note that Models Two (Figure 5) and Three (Figure 6, discussed later) could have been combined to form a saturated model, however the restrictive sample size



Note: T1 = teamwork; T2 = customer focus; T3 = oral expression; T4 = tolerance; T5 = comprehension. The observed variables 't1' though to 't5' represent trait judgements that correspond to each associated latent trait.

Figure 5. Model Two: Dimension-Specific CFA Model.

disallowed this. Figure 5 shows a graphical representation of Model Two. Latent variables reflected trait-based dimensions, and observed variables reflected trait-based judgements made across the exercises in the AC. Table 44 shows the standardised factor loadings from the dimension-specific model. These look promising on initial inspection, however, the goodness-of-fit indices shown in Table 45 indicate a poor fit for the dimension-specific model (Byrne, 2001). The number of parameters in Model Two totalled 40, which is near the minimum number, relative to sample size, suggested by Bentler and Chou (1987).

Table 44

Standardised Factor Loadings for Model Two: The Dimension-Specific CFA Model

Trait Judgements	Dimensions				
	T1	T2	T3	T4	T5
Ex 1, t1	.54				
Ex 2, t1	.53				
Ex 3, t1	.67				
Ex 1, t2		.61			
Ex 2, t2		.60			
Ex 3, t2		.70			
Ex 1, t3			.46		
Ex 2, t3			.55		
Ex 3, t3			.60		
Ex 1, t4				.58	
Ex 2, t4				.66	
Ex 3, t4				.70	
Ex 1, t5					.52
Ex 2, t5					.61
Ex 3, t5					.69

Note: T1 = teamwork; T2 = customer focus; T3 = oral expression; T4 = tolerance; T5 = comprehension. Ex1 = approach exercise; Ex2 = closing exercise; Ex3 = returns exercise. The observed variables 't1' though to 't5' represent trait judgements made in each exercise and corresponding to each associated latent trait.

Table 45

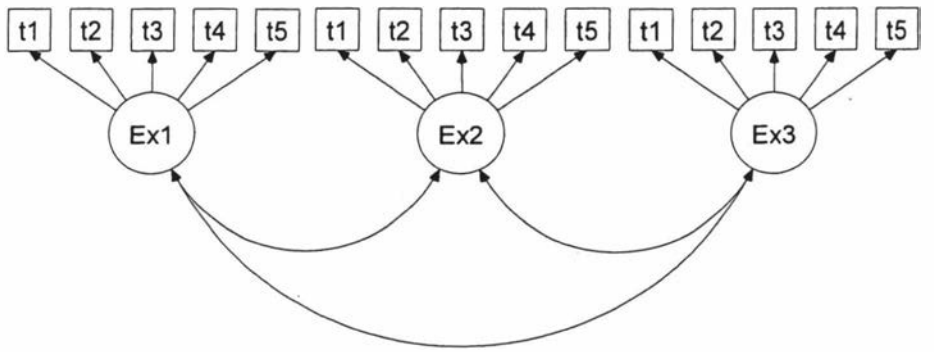
Selected Goodness-Of-Fit Indices for Model Two: The Dimension-Specific CFA Model

Index	Point Estimate
GFI	.524
AGFI	.286
CFI	.601
TLI	.476
RMSEA*	.247

*90% Confidence Interval for RMSEA (.233 < RMSEA < .261)

Model Three: The Exercise Effect CFA Model

The third model reflected the effect of different traits correlating highly within exercises (i.e., the exercise effect). Figure 6 shows a graphical representation of this model. In this case, latent variables represent heterotrait-monomethod correlations, and observed variables reflect trait judgements. Table 46 shows the standardised parameter estimates for the model represented in Figure 6.



Note: Ex1 = approach exercise; Ex2 = closing exercise; Ex3 = returns exercise. The observed variables 't1' though to 't5' represent trait judgements made in each exercise and corresponding to each associated latent exercise, where: t1 = teamwork; t2 = customer focus; t3 = oral expression; t4 = tolerance; t5 = comprehension.

Figure 6. Model Three: The Exercise Effect CFA Model.

Table 46

Standardised Factor Loadings for Model Three: The Exercise Effect CFA Model

Within Exercise Judgements	Exercises		
	Ex1	Ex2	Ex3
Ex 1, t1	.74		
Ex 1, t2	.84		
Ex 1, t3	.83		
Ex 1, t4	.87		
Ex 1, t5	.81		
Ex 2, t1		.73	
Ex 2, t2		.83	
Ex 2, t3		.83	
Ex 2, t4		.92	
Ex 2, t5		.88	
Ex 3, t1			.82
Ex 3, t2			.86
Ex 3, t3			.82
Ex 3, t4			.93
Ex 3, t5			.89

Note: Ex1 = approach exercise; Ex2 = closing exercise; Ex3 = returns exercise. The observed variables 't1' though to 't5' represent trait judgements made in each exercise and corresponding to each associated latent exercise, where: t1 = teamwork; t2 = customer focus; t3 = oral expression; t4 = tolerance; t5 = comprehension.

Table 46 shows relatively clear factor loadings for traits on exercises, consistent with the generalizability analysis. Table 47 shows selected goodness-of-fit indices for Model Three. Overall, the indices presented here suggested a mediocre fit between the proposed model and the observed data. The number of parameters relative to the number of cases was within the limits suggested by Bentler and Chou (1987).

Table 47

Selected Goodness-Of-Fit Indices for Model Three: The Exercise Effect CFA Model

Index	Point Estimate
GFI	.877
AGFI	.830
CFI	.946
TLI	.934
RMSEA*	.087

*90% Confidence Interval for RMSEA (.072 < RMSEA < .102)

The reader is cautioned about sample characteristics in the study above. Byrne (2001) and Raykov and Marcoulides (2000) note that normality assumptions are integral to SEM procedures, and are often neglected in practice. To gain some idea of sample characteristics, indices of univariate skewness and kurtosis were calculated for the abridged task-specific and dimension-specific data (see Cramer, 1994 for a discussion on how these indices are calculated). For the abridged task-specific data, skewness values ranged from –1.339 to 0.119, with a mean value of -0.679 (standard error = 0.178). Using the mean value as an estimate of overall skewness, the task-specific data were found to be significantly asymmetrical (z = -3.815, p < .001, 2-tailed), and thus, positively skewed. Kurtosis ranged from –1.009 to 2.153, with a mean value of –0.036 (standard error = 0.354). Using the mean value as an estimate of overall kurtosis, a significance test revealed that the data were not significantly platykurtic (where there are too few cases at the centre of a distribution), and therefore suggested evidence in favour of normality (z = -0.102, *ns*, 2-tailed). For the dimension-specific data, univariate skewness values ranged from –1.026 to –0.038, with a mean value of –0.059 (standard error = .178). Using the mean value as an

estimate of overall skewness, the dimension-specific data were not found to be significantly asymmetrical (z = -0.331, *ns*, 2-tailed). Kurtosis for the dimension-specific data ranged from -0.818 to 1.119, with a mean value of –0.187 (standard error = 0.354). Using the mean value as an estimate of overall kurtosis, a significance test revealed that the data were not significantly platykurtic, and therefore suggested evidence in favour of normality (z = -0.528, *ns*, 2-tailed). The deviations against normality in these data were therefore not deemed to be catastrophic. The task-specific data were found, on average, to be significantly positively skewed, however there was no salient evidence to suggest that kurtosis was problematic in these data.

The reader is cautioned that both the task-specific and the dimension-specific data sets failed to meet the assumptions of multivariate normality. Multivariate normality was assessed using Mardia's (1970) coefficient for the task-specific data = 22.757, and the dimension-specific data = 22.657. Values of 1.96 or less indicate non-significant multivariate kurtosis. Byrne (2001) asserts that in SEM, deviations from multivariate normality can lead to spuriously large χ^2 values, modest underestimation with respect to fit indices (particularly the TLI and the CFI), and spuriously low standard errors which may render spuriously significant regression paths in structural models. Note, "in practice, most data fail to meet the assumption of multivariate normality" (Byrne, 2001, p. 268).

Discussion

Generalizability Study

Study Three generally shows evidence in support of Hypothesis Two. Note that the task-specific and dimension-specific approaches in the study were compared in a

repeated measures design so as to hold raters, participants and assessment content constant. As can be seen in Table 34, clear exercise effects were found across both types of AC. This was evidenced by the *px* interaction (Kane, 1982; Kraiger & Teachout, 1990; Lievens, 2001b), which explained 30.7% of the variance in the task-specific model, and 33.6% of the variation in the dimension-specific model. Clearly under the dimension-specific model, this comparatively high source of variation makes little conceptual sense, and generally reflects a lack of evidence in favour of convergent and discriminant validity. Under a trait paradigm, one expects to measure variables that will endure in a relatively stable fashion across different situations. In the light of the previous discussion, even under a trait-paradigm, some variation across exercises might therefore be expected. However, in the dimension-specific approach, *px* was the greatest contributor to variance in scores. Note that under a task-specific model, the finding of a large amount of variation being attributable to *px* does indeed make conceptual sense. A detection of the profound effect of the situation and its influence on behaviour is considered integral and adaptive under the task-specific paradigm (Hartman, Roper & Bradford, 1979).

As with most forms of assessment in the selection context, the focus is on person variation across the various facets, because of the notion that assessment procedures of this nature aim to differentiate among people for decision purposes. Therefore, the principal focus in the present study concerns interactions between persons and facets and variance component estimates for the object of measurement. As mentioned, of particular interest in the present study is the interaction term *px* for both types of AC (Kane, 1982; Kraiger & Teachout, 1990; Lievens, 2001b). Lievens (2001a; 2001b) suggests that the interaction term *pd*, in the dimension-specific AC, reflects the extent to which dimensions (as a set) are useful for discriminating between

persons, that is, *pd* represents the extent to which the procedure holds a form of discriminant utility under a traditional trait paradigm. In Study Three this interaction term explained comparatively little of the variation in scores, at 1.8%. Thus, further evidence was found suggesting that the dimension-specific approach was not measuring trait-based variables because the *pd* interaction implies that dimensions were not comparatively useful for making differentiations among people.

Person variation is an important source of variance that needs to be given attention in ACs. An AC process must be efficacious in discriminating among people for decision purposes. The effect for the object of measurement, *p*, for the task-specific approach explained 27.4% of the total variance. The object of measurement, *p*, for the dimension specific approach was marginally higher, and explained 31.9% of the variance in scores. Note the slightly wider confidence intervals for this effect in Figures 2 and 3, suggesting some level of uncertainly in this variance estimate. The propensity for distinguishing among people for the two processes remains at a comparable level within the bounds of the respective confidence intervals for these person effects across the task-specific and dimension-specific processes. The reasons for these processes being able to discriminate among people in this way remains conceptually challenging for the dimension-specific approach, and conceptually comfortable for the task-specific approach, as evidenced in the high *px* interactions across the two approaches, and the low *pd* interaction in the dimension-specific approach.

Of particular interest are the similarities in patterning across the two types of AC. This is best seen in Figures 2 and 3 where similarities between the task-specific and dimension-specific approaches can be easily compared. Starting with the effects for *p*, *x*, *i:x* and their dimension-specific counterparts *d*, and *px* across ACs. The

similarities among analogous contributors to variance in the two Figures suggests that perhaps both ACs are isomorphic, or are at least similar, in their measurement outcomes. The major difference between the two models is that the task-specific approach makes conceptual sense, while the dimension-specific approach does not, as detailed earlier. Speaking speculatively, it is possible that the managerial assessors in this study are indeed treating the exercises in both of the AC models as stand alone work samples of situationally specific behaviour. This would be at odds with any form of trait-based measurement in ACs. Figures 2 and 3 suggest that credence may be given to most of the variance estimates in Study Three, apart from that for the exercise facet alone. The uncertain estimate of variance for the main effect for exercises suggests that it is difficult to draw conclusions with respect to this effect.

It should be noted by the reader that, given the results of the G study, the use of $E\rho^2_{Rel}$ and ϕ in this context is somewhat debateable. It is argued in the original monograph on G theory "While it is not assumed that *p* [the variance attributable to the object of measurement] is completely stable during the period to which the universe definition applies, it is taken for granted that *p*'s characteristics fluctuate around a typical value" (Cronbach et al., 1972, p. 363). That is to say, there is at least some stability of responding assumed when employing G and Phi. The use of these coefficients is perhaps questionable because the evidence from the G study suggests, in line with previous research, that the AC ratings reflect situationally specific responses, rather than stable characteristics. However, Cronbach et al. suggest that when the occasions of assessment are considered as samples of behaviour, it is "mathematically sound to define the universe score as the average over the time span [over which behavioural measurements will be made]" (p. 363). This could reflect overall performance on the exercises as samples of behavioural performance; a

conception that seems acceptable in the role of task-specific ACs where it is necessary to pool results at the end of the process to provide a summary rating for selection purposes (Lowry, 1997).

The calculation of G and Phi as indices of dependability in the context of a situationally specific form of assessment are possibly justified under the arguments above. It is interesting to note the similarities between the results of the two procedures. For the task-specific model, Table 35 shows that for relative decisions, $E\rho^2_{Rel}$ was calculated at 0.72, and for absolute decisions, ϕ was calculated at 0.70. These estimates were marginally higher than those calculated for the dimension-specific model in Table 36, where $E\rho^2_{Rel}$ was calculated at 0.71, ϕ was calculated at 0.68. Thus, the task-specific AC was found to be a marginally more dependable form of assessment than the dimension-specific model.

The interpretation of $E\rho^2_{Rel}$ and ϕ necessitates some deliberation at this point. While Lievens (2001a) cites Marcoulides (1989) and states that "Values equal or above .80 are considered to be acceptable" (Lievens, 2001a, p. 260), Marcoulides (1989) actually sets no such strict criterion for the interpretation of these coefficients. Indeed, Marcoulides has commented that he does not necessarily agree with such steadfast criteria for these indices (G. A. Marcoulides, personal communication, November 23rd, 2002). $E\rho^2_{Rel}$ and ϕ are Decision study (D study) values that should ideally be viewed in terms of the extent to which they increase relative to the costs associated with changing aspects of the facets of measurement in a particular model, for example changing the number of items or the number of dimensions. These coefficients can be examined by a researcher for the sole purpose of investigation into the values associated with a particular G study, rather than exclusively with comprehensive D studies, which look at the effects of changing the number of levels

of particular facets so as to determine effects on dependability. As such, the use of $E\rho^2_{Rel}$ and ϕ in this context is acceptable, and has been employed successfully in research on ACs (Arthur et al, 2000; Lievens 2001a). Shavelson and Webb (1991) suggest that $E\rho^2_{Rel}$ and ϕ are analogous to reliability coefficients in classical test theory. As a very general idea of the criteria for acceptability in cases such as those in the present study, it is probably more accurate to follow Shavelson and Webb (1991) than to follow Lievens (2001a) in the interpretation of $E\rho^2_{Rel}$ and ϕ. As summarised by Aiken (2003), the acceptability of a reliability coefficient can lie anywhere from between .60 or .70 and upwards, depending on the use of the data. As a general heuristic, the higher the coefficient, the better.

Table 35 also shows the interrater reliability for the task-specific model, ICC 1,1, calculated as 0.93. Overall inter-rater agreement on the task-specific model was found to be higher than that obtained for the dimension-specific model, ICC 1,1 calculated as 0.82. This finding is congruent with Lowry (1995) who reported that task-specific ACs yielded interrater reliability coefficients exceeding .80.

Factor Analysis

The factor analyses provided further evidence in favour of Hypothesis Two, and reinforced the findings in the G study. Table 38 shows the varimax rotated factor matrix for the task-specific AC. Table 40 shows the direct oblimin rotated factor matrix. Goodness of fit was reasonable for the three-factor solution, which accounted for 64.2% of the variance in the variables. Most of the communalities suggested that the variables were, by and large, well embedded within the factor structure. Item 6 on the Approach exercise could be regarded as an exception to this, with a comparatively low communality at .39. Only 17% of the residuals in the reproduced correlation

matrix were greater than .05 in absolute terms. As stated earlier, a cut-off of .4 was selected for noteworthy factor loadings. Given this criterion, relatively clear loadings of variables on exercises were evident. This is in congruence with the theoretical expectations of task-specific ACs, which consider exercises to act as stand-alone work samples of situationally specific performance.

Table 39 shows the varimax rotated five-factor solution for the dimension-specific AC. Table 41 shows the direct oblimin rotation. On initial inspection, the five-factor model accounts for a sizable amount of the variance in the variables at 76.5%. Communalities suggest that all of the variables are reasonably well embedded in the overall factor structure. Only one of the residuals in the reproduced correlation matrix had a value greater than .05 in absolute terms. Where goodness of fit appears to be promising on the surface, the evidence suggests that the factor structure of the ratings is conceptually problematic. Given a cut-off value of .4 for notable factor loadings, all of the variables load clearly onto three, as opposed to five, factors in Table 39. The exception to this is the variable 'comprehension' measured in the approach exercise, which bleeds across two factors. Aside from this, the factor loadings, clearly interpretable as the three exercises, are relatively clean. Generally, the fourth and fifth factors are redundant. This finding is typical of the heavily deliberated exercise effect seen in ACs. Different traits correlated highly within exercises, and same traits barely correlated across exercises. Thus, the dimension-specific AC displayed poor discriminant and convergent validity, when viewed from the traditional trait-based paradigm under which these processes operate. This exercise effect is less clear in the direct oblimin rotated pattern matrix shown in Table 41. This said, only three variables bleed across factors, and there is still a tendency

for variables to load onto exercises. Indeed, there is no clear evidence for trait-based measurement in Table 41.

These results of the factor analyses were congruent with those in the G study. In a dimension-specific AC, the intention is to measure trait-based variables under the trait-paradigm. This is why behaviours are classified under headings such as 'Comprehension' or 'Oral Expression', which are often referred to as 'competencies' or 'dimensions'. The reality is, no matter how they are termed, there is a trait-based expectation that raters will find some cross-situational patterning in behaviour. In ACs, this translates into a set of identical trait judgements, which should theoretically correlate highly across different exercises. However, the analysis in this, and other, studies suggests method variance in the dimension-specific AC. This finding does not correspond with the hypothetical expectation of the dimension-specific AC, thus, it makes little conceptual sense in that context.

Turning to the alternative task-specific paradigm, one treats each exercise as a stand-alone work sample of behaviour. No inference of stable traits is ever made. Thus, one would expect to obtain high correlations between the different behavioural items within an exercise under this paradigm. Where high factor loadings on exercises are problematic for the trait paradigm, for the behavioural paradigm, however, they are conceptually expected, adaptive, and admissible. Under the behavioural paradigm, high factor loadings on exercises reflect true variance in terms of situational specificity in behavioural responses.

Confirmatory Factor Analysis

The results of the CFA added emphasis to the results found in the previous analyses. The dimension-specific model (Model Two, Figure 5) emerged as the

poorest fitting model in the analysis (see Table 45). Model Three (see Figure 6), specifically investigated the extent to which heterotrait-monomethod correlations fitted the data. Overall, the goodness-of-fit indices indicated a mediocre fit for the exercise effect model (see Table 47).

The alternative to the dimension-specific models (Models Two and Three) was the task-specific model, Model One. An abridged version of this model was derived due to the restrictive sample size. Factor loadings for the task-specific model were high, and were consistent with the previous analyses on these data (see Table 42). Overall, the goodness-of-fit indices shown in Table 43 indicated a reasonable fit for the task-specific model, in line with the suggestions of Byrne (2001). Comparatively, the task-specific model was the best fitting of the three models tested.

Considerations

First and foremost, it could be argued that the present study employed a repeated measures design with no form of matching or counterbalancing the order of conditions (that is, the presentation of a task-specific followed temporally by a dimension specific approach). Thus, the order in which these conditions were presented may have affected the results obtained to some degree. However, there was only one logical order in which the conditions under study could be directed. Under a process such as an AC, in order to make a trait-based judgement of an individual, one must first witness a behavioural manifestation of that trait. This behavioural manifestation is then followed by a trait-based judgement. The reverse contingency cannot, and does not in practice, logically apply, as an assessor can neither reasonably nor defensibly make a behavioural judgement on the basis of a trait assumed to exist prior to the behavioural evidence. Because the initial step in AC methodology is to

document behaviours, a behavioural assessment is a natural consequence of having observed behavioural responses. The following natural progression in AC methodology and practice is to categorise these behaviours into a class of related behaviours. Thus, a behavioural assessment followed by a dimensional assessment is the natural order of events that transpires in an AC.

In addition to the above argument, it should be noted that, overall, more attention was given to the measurement of trait-based variables in this AC than to the measurement of behaviours in exercises. The behavioural checklists in the task-specific component of the AC displayed specific dimensions that were associated with each behavioural item. Thus, these checklists could be viewed as acting to maximise the possibility of trait measurement. The literature on ACs suggests that the presence of behavioural checklists should act to maximise conditions for trait measurement (Lievens, 1998). Training in the present study focused primarily on behaviour as a manifestation of trait variables. While the present study attempted to facilitate trait measurement in this regard, further credence could, perhaps, be given to the evidence in favour of a behavioural assessment as opposed to a trait-based assessment in this process. In a similar vein, the exercises employed in this study were of a relatively similar format. This design feature was intended to facilitate the manifestation of trait variables. The results would suggest that relatively minor fluctuations across exercises have an effect on behaviour, in line with Michel's theory (Michel, 1984).

The FOR procedure employed focused on the manifestation of dimensions only. Future research should look into whether FOR training, targeted at agreement in the ratings of task-specific measurements, could assist in improving the task-specific measurement model. The focus in task-specific AC training should shift away from a focus on trait manifestations *across* exercises, and should concentrate on behavioural

performance *within* an exercise itself. This approach may facilitate the measurement accuracy of the task-specific AC.

Consideration in this study should also be given to the restrictive sample size and the entry-level position under scrutiny, which may limit the level of ecological validity that the study might hold. Nevertheless, it is argued that 187 participants is a large group for an AC process, which tend to use much smaller numbers of people on a given assessment occasion (Ballantyne & Povah, 1995). ACs are often used for the selection of managerial personnel (Woodruffe, 1993). The generality of the above findings to higher-level positions cannot be definitively ascertained from the results of this study. Generality in this regard is suggested as a route for future research. The set of dimensions and the set of behavioural responses used in this study also require further research on different dimensions and behavioural responses to ensure generality across samples.

In the CFA, consideration also needs to be given to restrictive sample sizes when employing SEM analyses. Of import in SEM is the number of cases relative to the number of parameters estimated in a given model. As previously mentioned, Bentler and Chou (1987) suggested that at a minimum of five cases per parameter should be included in a given study. Byrne (2001) suggests that in small samples, goodness-of-fit indices (particularly the RMSEA and the TLI) can underestimate the true fit of a model. Small case numbers relative to the number of parameters estimated afflicted the full task-specific Model One. Therefore, an abridged version of the task-specific model was derived by randomly selecting items to create a smaller subset per exercise. The reader is therefore cautioned about possible limitations in the generality of this structural model to the entire task-specific data set. Sample size

restrictions rendered impossible the analysis of a fully saturated model that incorporated both exercise effects and dimensions effects (as in Arthur et al., 2000).

Theoretical Implications

The findings of Study Three are suggestive of a redefinition of the paradigm under which ACs currently operate. The suggestion is made that a task-specific paradigm may be more appropriate and theoretically justified than its dimension-specific counterpart. A multitude of past studies on ACs have viewed exercise effects as being indicative of halo effects, method effects, or measurement error (Carrick & Williams, 1999; Hough & Oswald, 2000; Schmidt & Ones, 1992). The present study viewed such effects as indicative that AC architects may have applied an inappropriate paradigm to a particular measurement instrument, thereby creating expectations that have not been upheld in the data on ACs to date. The exercise effect commonly observed in ACs appears to support this contention (Chan, 1996; Hough & Oswald, 2000; Schmidt & Ones, 1992).

The findings of the present study suggest that not only did the task-specific AC tend to produce ratings that made more sense psychometrically, the task-specific ratings also tended to be somewhat more dependable and reliable than the dimension-specific process. Such psychometric advantages imply that AC ratings can potentially become more useful to practitioners. Employment decisions related to development, selection and/or promotion based on AC ratings are more likely to be precise. The fairness with which such decisions are made under a task-specific process is more likely to be reinforced and justified over and above the dimension-specific process. Feedback on the basis of ratings that are anchored to specific tasks, rather than to nebulous dimensions, are more likely to lead to greater behavioural change (Thornton,

et al., 1995) in task-specific development centres, as opposed to the traditional dimension-specific approach.

Moreover, the task-specific approach may be more justifiable in court cases relating to employment decisions. Specific behavioural anchors specifying job related behaviours could reasonably be presented as a justification for employment decisions. Such information presents a less nebulous view of a person than does a trait-related assessment. Such suggestions should not be taken lightly as very few companies internationally investigate the extent to which their dimension-specific ACs are measuring constructs as intended (Spychalski, et al., 1997). A reasonable take on the literature would suggest that if a company is employing managers as assessors, which most do (Lowry, 1996; Muchinsky, 2000; Spychalski, et al., 1997), then the likelihood is that their AC will yield poor evidence of construct validity (Hough & Oswald, 2000; Schmidt & Ones, 1992), and therefore may be difficult to justify in court (Lowry, 1996; Norton, 1977). Not only is this important from a legal perspective, but it appears unethical to provide data for people on the basis of a model that is not psychometrically supported.

Given concerns about the cognitive load upon assessors in ACs (Lievens & Klimoski, 2001), and the limitations of managers as trait-based raters (Sagie & Magnezy, 1997), the task-specific approach to AC design possibly presents a straightforward treatment for problems associated with cognitive load and non-psychologist assessor panels. The very notion of finding classifications for behaviours under trait classes presents a highly complex task to a group of assessors who, primarily in practice, are not trained as psychological experts (Lowry, 1996; Muchinsky, 2000; Spychalski, et al., 1997). No such classification is necessary under

a task-specific approach. Thus, cognitive load upon assessors is, by design, also likely to be minimised under a task-specific model.

In the CFA, an abridged version of the task-specific model yielded the best fit when compared to the dimension-based and exercise effect models. With respect to AC data and structural models, Arthur et al.'s (2000) study is worthy of note. Arthur et al. tested only one CFA model from their data, which consisted of a saturated model incorporating exercise and dimension effects. Overall, they found an excellent fit for their mixed dimension/exercise model. The reason that mixed models fit well may be because such an array of variables are entered into such models. The practical use of mixed models in AC contexts is, however, questionable. Utilising the effects of monotrait-heteromethod and heterotrait-monomethod correlations for decision purposes appears overly burdensome. Also, there are currently no guidelines to show which effect (i.e., exercises or dimensions) should be given more or less weighting, other than the literature on exercise effects (heterotrait-monomethod correlations) commonly found in ACs (Hough & Oswald, 2000). Additionally, under the trait paradigm, the notion of relying on heterotrait-monomethod correlations to make decisions about people at all remains uncomfortable, and conceptually difficult to justify. In this study, the results pertaining to Models Two and Three suggest that heterotrait-monomethod correlations should be given more weighting than monotrait-heteromethod correlations.

The abridged task-specific Model One emerged as a reasonable fit, and was the best fitting of the three models tested. This may be considered encouraging in terms of a potentially practical model for AC evaluation methodology. Also, despite the fact that training did not focus on effects within exercises, both of the exercise centred models (Model One and Model Three) emerged as better fits than the

dimension based model. The dimension-based model on which AC training was focused emerged as a poor fit (Model Two). Thus, it would appear that more investigation is required into ACs that investigate exercise-centred performance. The most conceptually sound of the two models that concentrated on exercise performance would appear to be the task-specific AC. As argued elsewhere in this thesis, future research should look at methods to refine this approach to obtain a practical tool that could be used for reasonable employment decisions.

Chapter Five: General Discussion

A well-documented quandary in the AC literature is the lack of propensity for AC ratings to display the measurement of the trait-based variables they are intended to measure (Bycio et al., 1987; Carrick & Williams, 1999; Chan, 1996; Fleenor, 1996; Jones et al., 1991; Joyce et al., 1994; Russell, 1987; Lievens, 2002; Robertson et al., 1987; Silverman et al., 1986; Spector, 2000; Turnage & Muchinsky, 1982; Turnage & Muchinsky, 1984). The enigmatic nature of this finding is compounded by the notion that ACs tend to predict certain criteria, particularly related to promotion, yet the reasons for this predictive utility remain unidentified (Chan, 1996). The present set of three studies attempted to find new ways of interpreting the AC puzzle by investigating the notions of unintended latent trait measurement (Study One), a preliminary investigation into the assessment perceptions held, particularly by assessors (Study Two), and the primary study; an alternative to the prevailing paradigm underlying AC assessment (Study Three).

Study One generally found evidence against the contention that latent traits are unintentionally measured in ACs. Only one of the variables studied, tacit knowledge, in one of the two samples could cogently be argued as a meaningful *theoretical* predictor of OARs. This variable was the lone significant contributor to variance in a model that explained only 16% of the variance in OARs. The measurement of tacit knowledge in this sample was also found to be unreliable, making it difficult to ascertain the unified nature of the construct. Additionally, 16% of the variance associated primarily with tacit knowledge does not paint a particularly convincing picture as to the notions underlying AC measurement. As only 16% of the variance in OARs was explained in one sample, it would seem unlikely that the composite model

of self-efficacy, self-monitoring and tacit-knowledge would constitute a reasonable substitute for OARs. Another sample in Study One did not find any meaningful relationship between the composite model and OARs, showing further evidence that is unlikely that these variables act as the primary contributors to AC validity.

From another perspective, it could be argued that 16% of the variance explained in scores could be construed as fairly sizeable, given that the correlation was with a construct that was not intentionally measured in the AC. However, it is also a sizable inferential leap and probably incorrect to suggest, on the basis of this finding, that managers are tapping into managerial intelligence during the AC proce Given the poor record that ACs hold for measuring trait-based variables, this appear questionable. In any case, no matter how this relationship is construed, 16% of the variance explained in OARs is not a cogent enough explanation to warrant a replacement of the AC with a paper test of managerial tacit knowledge. Moreover, when correcting for validity shrinkage when generalising across different samples, tl variance in OARs explained by this composite dropped to around 1%. This could suggest that the composite external measures are not implicated in OAR derivation generally. Note, however, that managerial tacit knowledge has been found to be unrelated to traditional intelligence (IQ) test scores (Wagner & Sternberg, 1991). Cook (1998) reports findings that suggest IQ scores relate to OARs. Thus, the combination of managerial tacit knowledge and IQ might yield more substantial level of relationship with OARs in certain samples that have strong requirements for managerial tacit knowledge. While these relationships potentially hold interest, nothing in Study One suggested that the relationships between self-efficacy, self-monitoring and tacit-knowledge with OARs would definitively explain what it is that ACs actually measure.

Study Three investigated the extent to which an alternative paradigm might assist in making sense of AC ratings. The suggestion that an alternative to the prevailing trait paradigm should be introduced into AC construction has been conveyed by a small faction of researchers (Gorham, 1978; Herriot, 1986; Klimoski & Brickner, 1987; Lowry, 1997; Robertson et al., 1987). These researchers, by and large, have suggested that treating AC exercises as stand-alone work samples of behaviour would be a more adaptive approach to the treatment of AC ratings, as research suggests that perhaps assessors treat AC exercises as behavioural samples anyway. No known research has compared the psychometric properties of a task-specific with that of a dimension-specific AC. Study Three sought to find some preliminary solutions to the question of construct validity in ACs by exploring the possibility that the ratings in ACs might reflect groups of situationally specific work samples.

In the AC in Study Three, it was found that exercise effects endured across the repeated measures task-specific and dimension-specific processes, as evidenced by strong *px* interactions and factor loadings on exercises. Various levels of other facets mirrored each other across the two processes, as can be seen across Figures 2 and 3. Also, the dimension-specific process showed a relatively low *pd* interaction, indicating that dimensions were not useful criteria for making decisions among candidates. This is of great concern, because ACs are frequently used to make decisions about the varying performances of different people and, in practice, these decisions are most commonly based on dimensions (Lowry, 1996; Sackett & Harris, 1988; Spychalski et al., 1997). Thus, when managers are employed as assessors, as is most commonly the case (Lowry, 1996; Muchinsky, 2000; Spychalski et al., 1997), there remains the likelihood that managers will not measure trait-based variables, as

evidenced in the comparatively large *px* interaction term. Thus, trait-based variab become conceptually problematic foundations for decision purposes in ACs.

The results of the G study in Study Three suggested that both the task-spec process and the dimension-specific process were useful for making distinctions among people, as evidenced by the similarly high component of variance for the object of measurement. To reiterate on the argument presented above, the proble that in a dimension-specific AC, decisions about people are likely to be made on t basis of dimensions, which evidently do not contribute a great deal to person variation. Person variation across the exercises themselves contributed a great dea more to variation in ratings across the dimension-specific and task-specific ACs ir Study Three. Therefore, performance on exercises possibly constitutes a more meaningful basis for decision purposes in this AC than dimensions do. The dimension-specific approach does not promote such bases for decisions under its t foundations. Thus it would seem that the task-specific model, which actively encourages person variation as a function of varying exercises, is worthy of future research concerning its practicability and generality across different samples.

A CFA added further evidence that dimensions were not useful criteria for decision making purposes, as the dimension-based model presented in Figure 5 emerged as a poor fit overall (see Table 45). The abridged task-specific model (se Table 43) emerged as a reasonable fit, and was the best fitting of the three models tested. The exercise effect model (see Table 47) emerged as a mediocre fit (accor to the guidelines summarised in Byrne, 2001). Future research with larger subject numbers will be necessary to verify these results, however the CFA gained promis evidence for the task-specific approach.

In New Zealand, the findings of Study Three present just as much concern as they do for the rest of the AC using world. In a recent newspaper article, top New Zealand consulting companies gave obvious credence to the trait-based nature of the ratings obtained in ACs (McCarthy, 2003). The comments made in this article implied that employment decisions were being made for people on the basis of their scores on competencies treated as trait-based categories. Comments were also made about ACs being useful developmentally in terms of contributing towards the improvement of an individual's skill base. There is a multitude of evidence to suggest that this is misleading, given the lack of support for the measurement of any relatively stable and enduring characteristic in a manager-assessed AC.

To elaborate on the findings in Study Three, there was no evidence to suggest the successful measurement of trait-based variables in the dimension-specific AC, as shown in the high *px* and low *pd* interaction terms, the relatively clear factor loadings on exercises, the poor fit of the dimension-based and the reasonable fit of the task-specific structural model. The results of the dimension-specific AC appear to mimic the patterns expressed in the task-specific approach (see Figures 2 and 3), suggesting that the two forms of assessment are measuring something isomorphic, or at least similar. The difference between the two approaches is that the psychometric patterns found in the dimension-specific AC make no clear conceptual sense, under the notion that the process was not measuring the trait-based categories it was intended to measure. Rather, the results are suggestive of the highly deliberated exercise effect found in ACs. Efforts were made to maximise the possibility of trait-measurement in this regard, with the employment of behavioural checklists displaying appropriate trait categories, the use of fewer dimensions to reduce cognitive load, the use of frame of reference training, and the use of exercises of a similar format. Under the task-

specific model, however, the psychometric properties do make conceptual sense, in that situationally specific responses were expected under this approach. Thus, the suggestion drawn from Study Three is that when the task-specific and the dimension-specific processes are used to measure the same behavioural output, the task-specific approach holds a stronger theoretical justification over the dimension-specific approach. The reader is warned, however, that further research is needed for the generality of these conclusions. Particularly, attention should be drawn to model effects that may be specific to this sample, for example the position being assessed, the set of dimensions, the set of behaviours, and the set of exercises employed.

Psychometrically, the task-specific model makes more conceptual sense than the dimension-specific approach in Study Three, and moreover, the task-specific approach yielded slightly greater dependability and inter-rater agreement than its dimension-specific counterpart (see Tables 35 and 36). Given these findings, it appears that the task-specific model of assessment may be more appropriate in the more common situation where managerial assessors are employed. Further research will be needed to confirm this suggestion. As discussed earlier, there may be other gains associated with the task-specific approach in addition to psychometric arguments, including, as detailed in the introduction, developmental feedback advantages (Adams, 1990; Mueller & Dweck, 1998), legal defensibility, ease of training, increased measurement precision, and the related potential for improvements to the AC process. These features could be aided with an understanding of what AC ratings actually mean. Table 48 details other advantages associated with the task-specific approach to AC design, relative to the traditional dimension-specific approach.

Table 48

Advantages of the Task-Specific Approach Relative to the Dimension-Specific Approach to AC Design

Task-Specific Design	Dimension-Specific Design
Can potentially use psychologist or non-psychologist assessors to yield construct evidence	Should ideally employ psychologist assessors to yield construct evidence – likely to incur greater costs as a result
Lower number of inferences as behavioural checklists are used as the primary data set for decisions	Higher number of inferences, as one extrapolates trait-based variables from behavioural checklists
Can potentially use very different exercises without undermining the validity of the assessment	Restricted to the use of very similar exercises only
Can assess 8-15 behavioural items per exercise	Should ideally assess 4-5 traits per exercise
Can assess different behaviours in each exercise	Should ideally repeat the measurement of a trait at least three times across exercises
Training is simplified by a focus on behaviours only	Training is complicated by trait extrapolations from behaviours
Less cognitive demands on assessors due to less complex inferences	More cognitive demands on assessors due to complex trait inferences
Evidence in this study suggests that construct valid ratings are obtained	A multitude of evidence suggests that construct valid ratings are not obtained
Less time consuming and therefore less costly, because there are fewer steps in the assessment process	More steps in the assessment process, therefore more time consuming and costly
Developmental feedback more likely to lead to adaptive behavioural change	Developmental feedback less likely to lead to adaptive behavioural change
Renders task-based training needs readily identifiable	Renders training needs in more vague, categorical terms
Situationally specific responses, consistent patterns of behaviour, and/or combinations of these are considered conceptually acceptable	Consistent patterns of behaviour under trait categories are considered acceptable
More likely to be justified in court because measurement intentions are more likely to be reflected in ratings	Less likely to be justifiable in court, because dimension-specific ACs have a history of psychometric problems

In relation to ACs that are commonly used in practice, attention should be drawn to the notion that there is some debate and confusion surrounding the intentio when using dimensions in dimension-specific ACs. Byham (1980) states that the dimension categories act as nominal classes only, and describes them as "a descriptic under which behaviour can be reliably classified" (p. 29). This definition could lead to confusion. If behaviours were to be classified under some form of nominal category, then surely one would expect reasonable correlations between the behaviours within a category label? The very definition of a category implies that its function is to provide a class or division for a subset of related elements. Even before Sackett and Dreher's (1982) seminal paper, Gorham (1978) was aware of the confusion that such categories might instigate, and suggested that dimensional categories in ACs should be abandoned completely. Indeed, the very notion of a categorical label may well lead assessors to expect a trait-based judgement (Sackett, 1987). This is because category titles probably promote the idea that decision makers should seek to make a judgement of characteristics that are relatively stable and enduring on the basis of behavioural elements that appear to be meaningfully related to one another. Such is the basis for trait categorisations that form their origins from observable behavioural responses, and by design, ACs have influenced, motivated, and encouraged such a classification.

Sackett (1987) deliberates on the intention of construct measurement in ACs, and states that the "ratings of a dimension across exercises aren't intended as merely repeated measures that should correlate perfectly" (p. 19). Instead, the intention, according to Sackett, is to measure partially overlapping behavioural samples across exercises. This said, Sackett argues further that if there are near zero correlations between the measurements of the same construct across different exercises, then the

notion of an overall score based on dimensions becomes problematic. This commonly appears to be the case with the widely deliberated exercise effect finding. A great deal of the literature to date appears to have focused on maximising the possibility that trait-based variables will be measured. Some of these methods have been imaginative, inventive, and even curious. However, as a body of literature, neither definitive nor completely cogent solutions to the AC enigma have been supplied from a trait-based perspective.

The results of Study Two suggested that the non-psychologist assessors in a nationally respected AC employed in Auckland, New Zealand, did not tend to differentiate the paradigm under which they were assessing. This might suggest that they were not aware of the way in which they should approach the assessment, or on which foundation they should base their assessment. These findings add colour to the picture presented by Sagie and Magnezy (1997) and Lievens and Conway (2001) who found that managerial ratings did not tend to reflect trait-based variables. It should be noted that even experienced clinical psychologists display limitations in the reliability of their assessment of individuals (Persons & Bertagnolli, 1999; Persons, Mooney & Padesky, 1995). These studies found that on some of the particular factors under scrutiny, clinical psychologists displayed moderate and even poor inter-rater agreement. It was also found that whether the psychologist held a Ph.D. was an important determinant of the level of accuracy in assessment. The expectation that managers should be able to perform an assessment of an individual on the basis of complex notions such as traits may be unrealistic, given this comparison. The use of I/O psychologists as assessors in ACs appears to be unrealistic with the associated cost, the absence of the psychosocial advantages associated with having managers assess their own staff, and the exclusion of job-specific/employer-specific subject

matter expert knowledge that managers possess (relative to external consultant I/O psychologists).

The focus on traits alone in the AC literature appears to be restrictive when comparisons are made with clinical assessment, which formed the very origins of psychological assessment in Western society (Anastasi & Urbina, 1997). Contemporary clinical approaches to assessment take a much more holistic view of behavioural responses, and acknowledge such factors as behaviour, physiological responses, cognitions, stimuli and situations, rather than merely focussing on trait-based categorisations alone (Bond, 1998). While the clinical approach to assessment rightfully attempts to tap a comprehensively rich source of information about a particular individual, it is probable that such an in-depth analysis is not necessary in the organisational arena. This said, the essence of the clinical form of assessment suggests that the contemporary approach is not to focus specifically on trait-based variables on which to base decisions, and that a more holistic view is necessitated. It is argued that, perhaps specifically for managerial assessors, a paradigm encompassing behavioural responses contingent on situations is appropriate. The resulting information is likely to provide a rich and useful assessment on which to base decisions. Such an assessment would, by design, acknowledge variation among individuals' behaviour, and the effect of the situation on that behaviour, rather than investigating the extent to which an individual varied on trait-based variables for which there is little empirical evidence in the AC context. The foundations of the task-specific assessment will form its bases on information that is more likely to be justifiable and assured.

Study Three demonstrated evidence in favour of a paradigm that rejects the use of category labels, and instead focuses on the operational definitions of behaviour

relative to situational contingencies. This study suggests that while a task-specific approach may make more conceptual sense psychometrically, it has the potential to increase the quality and decrease the ambiguity and subjectivity associated with developmental feedback given to employees. It also has the potential to increase the quality and precision with which selection decisions are made. The move to the task-specific approach is a radical leap from the existing dimension-specific paradigm. Some may argue that such an alternative is impractical because it will entail overly detailed job analyses and a tailored AC for each organisation. It is acknowledged that it is practically difficult to maintain such bespoke detail in real-world scenarios. However, it is argued that job analyses should always be a defining feature in the development of assessment programs so as to maintain job relevance and defensibility. While such analyses may not always be at the level of detail required for the construction of a task-specific AC, it is possible that taxonomies of tasks that relate to specific positions could be made available through an item bank. These could be applied in relation to a job analysis that would potentially require less task-related detail than the inductive approaches described in Lowry (1997).

Additionally, some practitioners may feel uncomfortable about discarding competency categories. In actual fact, the competency categories could still exist in the background in a task-specific AC, but would be treated as labels for groups of behaviours only. In practice, the very operational definitions of these categories would be applied in the AC. The major difference under the task-specific, when compared to the dimension-specific paradigm, would be the omission of any inference of stable traits. A person's performance on a given exercise would become the new unit of measurement, rather than the label attached to a set of behaviours said to underlie a given competency.

It will be desirable for future studies to investigate the generality of the findings in Study Three, due to the possibly sample-specific considerations detaile earlier. Also, further research into the predictive validity of the task-specific appro will be vital to ensure its worth as a tool for decision-making. While Study Three found preliminary evidence that the task-specific approach is conceptually sound, i remains silent on the notions surrounding whether this approach can explain simila a greater amounts of variation in criterion scores such as work performance or promotability. If a task-specific approach can explain variation in scores for these criteria, then the evidence in this study suggests that the reasons for this relationshi will ultimately be less of an enigma.

References

Aamodt, M. G. (1999). *Applied industrial/organizational psychology* (3rd ed.). CA: Wadsworth.

Adams, K. A. (1997). *The effect of the rating process on construct validity: Reexamination of the exercise effect in assessment center ratings.* Unpublished master's thesis, University of Houston, Houston, TX.

Adams, S. R. (1990). *Impact of assessment center method and categorization scheme on schema choice and observational, classification, and memory accuracy.* Unpublished doctoral dissertation, Colorado State University, Ft. Collins, CO.

Ahmed, Y., Payne, T. & Whiddett, S. (1997). A process for assessment exercise design: A model of best practice. *International Journal of Selection and Assessment, 5*(1), 62-68.

Aiken, L. R. (2003). *Psychological testing and assessment* (11th ed.). Boston: Allyn and Bacon.

Anastasi, A. & Urbina, S. (1997). *Psychological testing* (7th ed.). NJ: Prentice Hall.

Anderson, N., Silvester, J., Cunningham-Snell, N. & Haddleton, E. (1999). Relationships between candidate self-monitoring, perceived personality, and selection interview outcomes. *Human Relations, 52,* (9), 1115-1131.

Arkin, R. M. (1981). Self-presentational styles. In J. T. Tedeschi (Ed.), *Impression management theory and social psychological research* (pp 311-330). New York: Academic Press.

Arthur, W., Woehr, D. J. & Maldegen, R. (2000). Convergent and discriminant validity of assessment center dimensions: A conceptual and empirical re-examination of the assessment center construct-related validity paradox. *Journal of Management, 26*(4), 813-835.

Arvey, R. D. & Murphy, K. R. (1998). Performance evaluation in work settings. *Annual Review of Psychology, 49,* 141-168.

Asher, J. J. & Sciarrino, J. A. (1974). Realistic work sample tests: A review. *Personnel Psychology, 27*, 519-533.

Ballantyne, I. & Povah, N. (1995). *Assessment and development centres.* Hampshire: Gower.

Bandura, A. (1977). *Social learning theory.* Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1982). Self-efficacy mechanism in human agency. *American Psychologist, 37,* 122-147.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control.* New York: W. H. Freeman.

Baron, H. & Janman, K. (1996). Fairness in the assessment centre. In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology, Vol. 11,* (pp. 61-113). Chichester: John Wiley a Sons.

Barrett, P. (1996). *PsWin psychometric software suite for Windows* [Computer program]. University of Liverpool, Department of Psychology.

Barrick, M. R. & Mount, M. K. (1991). The big five dimensions and job performance: A meta-analysis. *Personnel Psychology, 44,* 1-26.

Baumeister, R. F. (1982). A self-presentational view of social phenomena. *Psychological Bulletin, 91,* 3-26.

Belsley, D. A., Kuh, E. & Welsch, R. E. (1980). *Regression Diagnostics.* John Wile and Sons: New York.

Bem, D. J. & Funder, D. C. (1978). Predicting more of the people more of the time: Assessing the personality of situations. *Psychological Review, 85*(6), 485-501

Bentler, P. M. & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods and Research, 16,* 78-117.

Bernardin, H. J. & Buckley, M. R. (1981). Strategies in rater training. *Academy of Management Review, 6,* 205-212.

Blanchard, P. N. & Thacker, J. W. (1998). *Effective training: Systems, strategies, an practices.* NJ: Prentice Hall.

Bobko, P. (1990). Multivariate correlational analysis. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology, Vol 1* (pp. 637-686). Palo Alto, CA: Consulting Psychologists Press.

Bobko, P. (2001). *Correlation and regression: Applications for industrial organizational psychology and management* (2nd ed.). Thousand Oaks, CA: Sage.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bond, F. W. (1998). Utilising case formulations in manual-based treatments. In M. Bruch & F. W. Bond, *Beyond diagnosis: Case formulation approaches in CBT* (pp. 185-206). Chichester: Wiley & Sons Ltd.

Borman, W. C. (1977). Consistency of rating accuracy and rater errors in the in the judgement of human performance. *Organizational Behavior and Human Performance, 20,* 238-252.

Bosscher, R. J. & Smit, J. H. (1998). Confirmatory factor analysis of the general self-efficacy scale. *Behaviour Research & Therapy, 36*(3), 339-343.

Brannick, M. T., Michaels, C. E. & Baker, D. P. (1989). Construct validity of in-basket scores. *Journal of Applied Psychology, 74,* 957-963.

Bray, D. W. & Grant, D. L. (1966). The assessment center in the measurement of potential for business development. *Psychological Monographs, 80*, 1-27.

Brennan, R. L. (2000). (Mis)conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14-20.

Brennan, R. L. (2001a). *Generalizability theory.* New York: Springer Verlag.

Brennan, R. L. (2001b). *Manual for urGENOVA.* Iowa City, IA: Iowa Testing Programs, University of Iowa.

Briggs, S. R., Cheek, J. M. & Buss, A. H. (1980). An analysis of the self-monitoring scale. *Journal of Personality and Social Psychology, 38*, 679-686.

Browne, M. W. & Cudeck, (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 445-455). Newbury Park, CA: Sage.

Buckner, M. (1984). *An evaluation of the effectiveness and reliability of a videotaped assessment center as compared to a live assessment center.* Unpublished doctoral dissertation, Georgia State University.

Bycio, P., Alvares, K. M. & Hahn, J. (1987). Situation specificity in assessment center ratings: A confirmatory factor analysis. *Journal of Applied Psychology, 72,* 463-474.

Byham, W. C. (1970). Assessment centers for spotting future managers. *Harvard Business Review, 48,* 150-160.

Byham, W. C. (1980, February). Starting an assessment center the right way. *Personnel Administrator,* 27-32.

Byrne, B. M. (2001). *Structural equation modeling with AMOS: Basic concepts, applications, and programming.* Mahwah, NJ: Lawrence Erlbaum Associates.

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Campion, J. E. (1972). Work sampling for personnel selection. *Journal of Applied Psychology, 56*(1), 40-44.

Carrick, P. & Williams, R. (1999). Development centres: A review of assumptions. *Human Resource Management Journal, 9*(2), 77-92.

Cascio, W. F. & Phillips, N. F. (1979). Performance testing: A rose among the thorns? *Personnel Psychology, 32,* 751-766.

Chan, D. (1996). Criterion and construct validation of an assessment centre. *Journal of Occupational and Organizational Psychology, 69,* 167-181.

Chatterjee, S., Hadi, A. S. & Price, B. (2000). *Regression analysis by example* (3rd ed.). New York: John Wiley & Sons.

Chow, S. L. (1996) *Statistical significance: Rationale, validity and utility*. CA: Sage Publications.

Cohen, B., Moses, J. L. & Byham, W. C. (1974). *The validity of assessment centers: A literature review*. Pittsburgh, PA: Development Dimensions Press.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences*. London: Lawrence Erlbaum Associates.

Colonia-Willner, R. (1998). Practical intelligence at work: Relationship between aging and cognitive efficiency among managers in a bank environment. *Psychology and Aging, 13,* 45-57.

Comrey, A. L. & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). New Jersey, Lawrence Erlbaum.

Cook, M. (1998). *Personnel selection: Adding value through people.* Chichester: John Wiley & Sons.

Coolican, H. (1999). *Research methods and statistics in psychology* (2nd ed.). London: Hodder & Stoughton.

Cramer, D. (1994). *Introducing statistics for social research: Step-by-step calculations and computer techniques using SPSS.* New York: Routledge.

Cronbach, L. J. (1955). Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin, 52,*177-193.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: John Wiley.

Cronbach, L. J. & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Day, D. V., Schleicher, D. J. & Unckless, A. L. (1996). *Self-monitoring and work-related outcomes: A meta-analysis.* Paper presented at the 11th Annual Conference of the Society for Industrial and Organizational Psychology, San Diego, CA.

Delprato, D. J. & Midgley, B. D. (1992). Some fundamentals of B. F. Skinner's behaviourism. *American Psychologist, 47,* 1507-1520.

Donahue, L. M., Truxillo, D. M., Cornwell, J. M. & Gerrity, M. J. (1997). Assessment center construct validity and behavioral checklists: Some additional findings. *Journal of Social Behaviour and Personality, 12,* 85-108.

Faul, F. & Erdfelder, E. (1992). GPOWER: A priori, post-hoc, and compromise power analyses for MS-DOS [Computer program]. Bonn, FRG: Bonn University, Department of Psychology.

Felson, R. B. (1981). An interactionist approach to aggression. In J. T. Tedeschi (Ed.), *Impression management theory and social psychological research* (pp 181-199). New York: Academic Press.

Feltham, R. T. (1989). Assessment centres. In P. Herriot (Ed.), *Handbook of assessment in organizations* (pp. 401-419). London: John Wiley & Sons.

Fleenor, J. W. (1996). Constructs and developmental assessment centres: Further troubling empirical findings. *Journal of Business and Psychology, 3,* 319-335.

Fletcher, C. & Anderson, N. (1998). A superficial assessment. *People Management, May,* 44-46.

Fletcher, C. & Kerslake, C. (1992). The impact of assessment centers and their outcomes on participants' self assessments. *Human Relations, 45,* 281-289.

Furnham, A., Crump, J. & Whelan, J. (1997). Validating the NEO Personality Inventory using assessor's ratings. *Personality and Individual Differences, 22,* 669-675.

Gabrenya, W. K., Jr. & Arkin, R. M. (1980). Self-monitoring scale: Factor structure and correlates. *Personality and Social Psychology Bulletin, 6,* 13-22.

Gaugler, B., Rosenthal, D., Thornton, G. & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72,* 493-511.

Gaugler, B. & Thornton, G. C. III (1989). Number of assessment dimensions as a determinant of assessor accuracy. *Journal of Applied Psychology, 74,* 611-618.

Goffin, R. D., Rothstein, M. G. & Johnston, N. G. (1996). Personality testing and t assessment center: Incremental validity for managerial selection. *Journal o Applied Psychology, 81,* 746-756.

Gold, M. S. & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation maximization. *Structural Equation Modeling, 7*(3), 319-355.

Gorham, W. A. (1978). Federal executive agency guidelines and their impact on th assessment center method. *Journal of Assesment Center Technology, 1*(1), 2 8.

Green, S. B. & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology 39,* 543-564.

Guilford, J. P. (1959). *Personality.* New York: McGraw-Hill.

Halman, F. & Fletcher, C. (2000). The impact of development centre participation and the role of individual differences in changing self-assessments. *Journal Occupational and Organizational Psychology, 73,* 423-442.

Handyside, J. & Duncan, C. (1954). Four years later on: A follow up of an experiment in selecting supervisors. *Occupational Psychology, 28,* 9-23.

Harris, M. M., Becker, A. S. & Smith, D. E. (1993). Does the assessment center scoring method affect the cross-situational consistency of ratings? *Journal of Applied Psychology, 78*(4), 675-678.

Hartmann, D. P., Roper, B. L. & Bradford, D. C. (1979). Some relationships between behavioral and traditional assessment. *Journal of Behavioral Assessment, 1*(1), 3-21.

Harvey, R. J. (1991). *Job analysis.* In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd ed.) (pp. 71-163). Palo Alto, CA: Consulting Psychologists Press.

Herriot, P. (1986). Assessment centres revisited. *Guidance and Assessment Review, 2*(3), 7-8.

Highhouse, S. & Harris, M. M. (1993). The measurement of assessment center situations: Bem's template matching technique for examining exercise similarity. *Journal of Applied Social Psychology, 23*(2), 140-155.

Hough, L. M. & Oswald, F. L. (2000). Personnel selection: Looking toward the future – remembering the past. *Annual Review of Psychology, 51,* 631-664.

Howard, A. (1997). A reassessment of assessment centers, challenges for the 21st century. *Journal of Social Behavior and Personality, 12,* 13-52.

Hunter, J. E. & Hunter, R. F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*(1), 72-98.

International Task Force on Assessment Center Guidelines. (2000). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management, 29,* 315-331.

Joiner, D. (2002). Assessment centers: What's new? *Public Personnel Management, 31*(2), 179-185.

Jones, A., Herriot, P., Long, B., & Drakeley, R. (1991). Attempting to improve the validity of a well-established assessment centre. *Journal of Occupational Psychology, 64,* 1-21.

Joyce, L. W., Thayer, P. W., & Pond, S. B. (1994). Managerial functions: An alternative to traditional assessment center dimensions. *Personnel Psychology, 47,* 109-121.

Kane, M. T. (1982). A sampling model for validity. *Applied Psychological Measurement, 6,* 125-160.

Kaplan, R. M. & Saccuzzo D. P. (2001). *Psychological Testing: Principles, applications, and issues* (5th ed.). Belmont, CA: Wadsworth.

Kenrick, D. T. & Funder, D. C. (1991). The person-situation debate: Do personality traits really exist? In N. J. Derlega, B. A. Winstead, & W. H. Jones (Eds.), *Personality: Contemporary theory and research* (pp. 150-174). Chicago: Nelson-Hall.

Kleinmann, M. (1993). Are rating dimensions in assessment centers transparent for participants? Consequences for criterion and construct validity. *Journal of Applied Psychology, 78,* 988-993.

Klem, L. (2000). Structural equation modeling. In L. G. Grimm, G. Laurence & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 227-260). Washington, DC: American Psychological Association.

Klimoski, R. J. & Brickner, M. (1987). Why do assessment centers work? The puzzle of assessment center validity. *Personnel Psychology, 40,* 243-260.

Klimoski, R. J. & Strickland, W. J. (1977). Assessment centers – valid or merely prescient. *Personnel Psychology, 30*, 353-361.

Kolb, J. A. (1998). The relationship between self-monitoring and leadership in student project groups. *Journal of Business Communication, 35*(2), 264-282.

Kraiger, K. & Teachout, M. S. (1990). Generalizability theory as construct-related evidence for the validity of job performance ratings. *Human Performance, 3,* 19-35.

Kudisch, J. D., Ladd, R. T. & Dobbins, G. H. (1997). New evidence on the construct validity of diagnostic assessment centers: The findings may not be so troubling after all. In R. E. Riggio & B. T. Mayes (Eds.), *Assessment centers: Research and applications* [Special Issue]. *Journal of Social Behavior and Personality, 12,* 129-144.

Lance, C. E., Newbolt, W. H., Gatewood, R. D., Foster, M. R., French, N. R. & Smith, D. (2000). Assessment center exercise factors represent cross-situational specificity, not method bias. *Human Performance, 13*(4), 323-353.

Lane, J. (1992). Methods of assessment. *Health Manpower Management, 18*(2), 4-6.

Lebreton, J. M., Binning, J. F. & Hesson-McInnis, M. S. (1998). *The effects of measurement structure on the validity of assessment center dimensions: The clinical-statistical debate revisited.* Paper presented at the Annual Meeting of the Academy of Management, Sa Diego, CA.

Lennox, R. D. & Wolfe, R. N. (1984). Revision of the self-monitoring scale. *Journal of Personality and Social Psychology, 46*, 1349-1364.

Licht, M. H. (1995). Multiple regression and correlation. In L. G. Grimm, G. Laurence & P. R. Yarnold (Eds.), *Reading and understanding multivariate statistics* (pp. 19-64). Washington, DC: American Psychological Association.

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6*, 141-152.

Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*(2), 255-264.

Lievens, F. (2001b). Assessors and use of assessment center dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior, 22,* 203-221.

Lievens, F. (2002). Trying to understand the different pieces of the construct validity puzzle of assessment centers: An examination of assessor and assessee effects. *Journal of Applied Psychology, 87*(4), 675-686.

Lievens, F. & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large-scale evaluation of multitrait-multimethod studies. *Journal of Applied Psychology, 86*(6), 1202-1222.

Lievens, F. & Klimoski, R. J. (2001). Understanding the assessment center process: Where are we now? In C.L. Cooper & I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology, Vol. 16,* (pp. 245-286). Chichester: John Wiley and Sons.

Lievens, F. & Van Keer, E. (2001). The construct validity of a Belgian assessment centre: A comparison of different models. *Journal of Occupational and Organizational Psychology, 74,* 373-378.

Lopez, F. M. (1988). Threshold traits analysis system. In S. Gael (Ed.), *The job analysis handbook for business, industry, and government* (Vol.2, pp. 880-901). New York: Wiley.

Lopez, F. M., Kesselman, G. A. & Lopez, F. E. (1981). An empirical test of a trait-oriented job analysis technique. *Personnel Psychology, 34,* 479-502.

Lowry, P. E. (1988). The assessment center: Pooling scores or arithmetic decision rule? *Public Personnel Management, 17*(1), 63-71.

Lowry, P. E. (1995). The assessment center process: Assessing leadership in the public sector. *Public Personnel Management, 24*(4), 443-450.

Lowry, P. E. (1996). A survey of the assessment center process in the public sector. *Public Personnel Management, 25*(3), 307-321.

Lowry, P. E. (1997). The assessment center process: New directions. In R.E. Riggio & B.T. Mayes (Eds.), Assessment centers: Research and applications [Special issue]. *Journal of Social Behavior and Personality, 12*(5), 53-62.

MacCallum, R. C., Browne, M. W. & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1,* 130-149.

Marcoulides, G. A. (1989). The application of generalizability analysis to observational studies. *Quality and Quantity, 23,* 115-127.

Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57,* 519-530.

Matthews, G. & Deary, I. J. (1998). *Personality traits.* Cambridge: Cambridge University Press.

McCarthy, A. (2003, January 20). Applicants grab chance to shine. *The New Zeala Herald,* p. E1.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits c our capacity for processing information. *Psychological Review, 63,* 81-97.

Mischel, W. (1968). *Personality and assessment.* New York: Wiley.

Mischel, W. (1973). Toward a cognitive social learning conceptualization of personality. *Psychological Review, 36,* 163-183.

Mischel, W. (1984). Convergences and challenges in the search for consistency. *American Psychologist, 39*(4), 351-364.

Moser, K., Diemand, A. & Schuler, H. (1996). Inconsistency and social skills as tw components of self-monitoring. *Diagnostica, 42,* 268-283.

Muchinsky, P. M. (2000). *Psychology applied to work* (6th ed.). Belmont, CA: Wadsworth.

Mueller, C. M. & Dweck, C. S. (1998). Praise for intelligence can undermine children's motivation and performance. *Journal of Personality and Social Psychology, 75*(1), 33-52.

Murphy, K. R. & Cleveland, J. N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks, CA: Sage Publications.

Murphy, K. R. & Davidshofer, (2001). *Psychological testing: Principles and applications.* (4th ed.). NJ: Prentice Hall.

Murphy, K. R. & Myors, B. (1998). *Statistical power analysis: A simple and genera model for traditional and modern hypothesis tests.* Mahwah, NJ: Lawrence Erlbaum Associates.

Neidig, R. D. & Neidig, P. J. (1984). Multiple assessment center exercises and job relatedness. *Journal of Applied Psychology, 69,* 182-186.

Norton, S. D. (1977). The empirical and content validity of assessment centers vs. traditional methods for predicting managerial success. *Academy of Management Review, 2,* 442-445.

Norton, S. D. (1981). The assessment center process and content validity: A reply to Dreher and Sackett. *Academy of Management Review, 6,* 561-566.

Nunnally, J. C. & Bernstein, I. H. (1994). *Psychometric Theory* (3rd ed.). New York: McGraw Hill.

O'Cass, A. (2000). A psychometric evaluation of a revised version of the Lennox and Wolfe revised self-monitoring scale. *Psychology & Marketing, 17*(5), 397-419.

Paton, D. & Jackson, D. J. R. (2002). Developing disaster management capability: An assessment centre approach. *Disaster Prevention and Management, 11*(2), 115-122.

Paulhus, D. L. (2002). Socially desirable responding: The evolution of a construct. In Braun, H. I. & Jackson, D. N. (Eds.), *The role of constructs in psychological and educational measurement* (pp. 49-69). Mahwah, NJ, Lawrence Erlbaum Associates

Persons, J. B. & Bertagnolli, A. (1999). Inter-rater reliability of cognitive-behavioral case formulations of depression: A replication. *Cognitive Therapy and Research, 23*(3), 271-283.

Persons, J. B., Mooney, K. A. & Padesky, C. A. (1995). Interrater reliability of cognitive-behavioral case formulations. *Cognitive Therapy and Research, 19*(1), 21-34.

Peterson, N. G. & Jeanneret, P. R. (1997). Job analysis: Overview and description of deductive methods. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp. 13-50). Palo Alto, CA: Davies-Black Publishing.

Pynes, J. & Bernardin, H. J. (1992). Mechanical vs. consensus-derived assessment center ratings: A comparison of job performance validities. *Public Personnel Management, 21,* 17-28.

Pynes, J., Bernardin, H. J., Benton, A. L. & McEvoy, G. M. (1988). Should assessment center dimension ratings be mechanically-derived? *Journal of Business and Psychology, 2*(3), 217-227.

Raykov, T. & Marcoulides, G. A. (2000). *A first course in structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Reilly, R. R., Henry, S. & Smither, J. W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology, 43,* 71-84.

Robertson, I. T., Gratton, L. & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology, 60,* 187-195.

Robertson, I. T. & Kandola, R. S. (1982). Work sample tests: Validity, adverse impact and applicant reaction. *Journal of Occupational Psychology, 55,* 171-183.

Robie, C., Adams, K. A. Osburn, H. G., Morris, M. A. & Etchegaray, J. M. (2000). Effects of the rating process on the construct validity of assessment center dimension evaluations. *Human Performance, 13*(4), 355-370.

Rosenthal, R. & Rosnow, R. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed.). New York: McGraw Hill.

Russell, C. J. (1987). Person characteristic versus role congruency explanations for assessment center ratings. *Academy of Management Journal, 30,* 817-826.

Russell, C. J., & Domm, D. R. (1995). Two field tests of an explanation of assessment centre validity. *Journal of Occupational and Organizational Psychology, 68,* 25-47.

Ryan, A., Daum, D., Bauman, T., Grisez, M., Mattimore, K., Nalodka, T. & McCormick, S. (1995). Direct, indirect and controlled observation and rating accuracy. *Journal of Applied Psychology, 80*(6), 664-670.

Sackett, P. R. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40,* 13-25.

Sackett, P. R. & Dreher, G. F. (1982). Constructs and assessment center dimensions: Some troubling empirical findings. *Journal of Applied Psychology, 67,* 401-410.

Sackett, P. R. & Dreher, G. F. (1984). Situation specificity of behavior and assessment center validation strategies: A rejoinder to Neidig and Neidig. *Journal of Applied Psychology, 69,* 187-190.

Sackett, P. L. & Hakel, M. D. (1979). Temporal stability and individual differences in using assessment center information from overall ratings. *Organizational Behavior and Human Performance, 23,* 120-137.

Sackett, P. R. & Harris, M. M. (1988). A further examination of the constructs underlying assessment center ratings. *Journal of Business and Psychology, 3*(2), 214-229.

Sadri, G. & Robertson, I. T. (1993). Self efficacy and work-related behaviour: A review and meta-analysis. *Applied Psychology: An International Review, 42*(2), 139-152.

Sagie, A. & Magnezy, R. (1997). Assessor type, number of distinguishable dimensions categories, and assessment centre construct validity. *Journal of Occupational and Organizational Psychology, 70,* 103-108.

Sanchez, J. I. & Frazer, S. L. (1992). On the choice of scales for task-analysis. *Journal of Applied Psychology, 77,* 545-553.

Sanchez, J. I. & Levine, E. L. (1989). Determining important tasks within jobs: A policy capturing approach. *Journal of Applied Psychology, 74,* 336-342.

Sartre, J. (1964). *Huis clos.* London: Methuen.

Saxe, R. & Weitz, B. A. (1982). The SOCO scale: A measure of the customer orientation of salespeople. *Journal of Marketing Research, 19,* 343-351.

Schleicher, D. J. (1999). A new 'frame' for frame of reference training: Enhancing the construct validity of assessment centers. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 60,* (1-B), 0384.

Schleicher, D. J. & Day, D. V. (1998). A cognitive evaluation of frame-of-reference rater training: Content and process issues. *Organizational Behavior and Human Decision-making Processes, 73*(1), 76-101.

Schleicher, D. J., Day, D. V., Mayes, B. T. & Riggio, R. E. (2002). A new frame for frame-of reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 87*(4), 735-746.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115-129.

Schmidt, F. L. & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1*(2), 199-223.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Schmidt, F. L., Ones, D. S. & Hunter, J. E. (1992). Personnel selection. *Annual Review of Psychology, 43,* 627-670.

Schmitt, N., Ford, J. K. & Stults, D. M. (1986). Changes in self-perceived ability as a function of performance in an assessment centre. *Journal of Occupational Psychology, 59,* 327-335.

Schmitt, N., Gooding, R., Noe, R. & Kirsch, M. (1984). Meta-analysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37,* 407-422.

Schmitt, N., & Ostroff, C. (1986). Operationalising the "behavioral consistency" approach: Selection test development based on a content-oriented strategy *Personnel Psychology, 39,* 91-108.

Schmitt, N., Schneider, J. & Cohen, S. (1990). Factors affecting validity of a regionally administered assessment center. *Personnel Psychology, 43,* 1-1

Schneider, J. & Schmitt, N. (1992). An exercise design approach to understandin assessment center dimension and exercise contructs. *Journal of Applied Psychology, 77,* 32-41.

Scholz, G. & Schuler, H. (1993). Das nomologische netzwerk des assessment centers: eine metaanalyse. *Zeitschrift fur Arbeits- und Organisationspsychologie, 37*(2), 73-85.

Seegers, J. (1997). What is an assessment centre? In P. Jansen & F. de Jongh (Eds.), *Assessment centres. A practical handbook.* Chichester: John Wiley and Sons.

Shavelson, R. J. & Webb, N.M. (1991). *Generalizability theory: A primer.* Newbu Park, CA: Sage Publications.

Sherer, M., Maddux, J. E. Mercandante, B., Prentice-Dunn, S., Jacobs, B. & Roge R. W. (1982). The self-efficacy scale: Construction and validation. *Psychological Reports, 51,* 663-671.

Shore, T. H., Thornton, G. C. III & Shore, L. M. (1990). Construct validity of two categories of assessment center ratings. *Personnel Psychology, 43,* 101-11

Shrout, P. E. & Fleiss, J. J. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Silverman, W. H., Dalessio, A., Woods, S. B. & Johnson, R. L. (1986). Influence of assessment center methods on assessors' ratings. *Personnel Psychology, 39,* 565-578.

Skinner, B. F. (1974). *About behaviorism.* New York: Knopf.

Snyder, M. (1974). Self-monitoring of expressive behavior. *Journal of Personality and Social Psychology, 30,* 526-537.

Snyder, M. (1987). *Public appearances / private realities: The psychology of self-monitoring.* New York: W. H. Freeman and Company.

Spector, P. E. (2000). *Industrial and organizational psychology: Research and practice* (2nd ed.). New York: Wiley.

Spencer, L. M. & Spencer, S. M. (1993). *Competence at work: Models for superior performance.* New York: Wiley.

Spychalski, A. C., Quinones, M. A., Gaugler, B. B. & Pohley, J. (1997). A survey of assessment center practices in organizations in the United States. *Personnel Psychology, 50,* 71-90.

Sternberg, R. J., Forsythe, G. B., Hedlund, J., Horvath, J. A., Wagner, R. K., Williams, W. M., Snook, S. A., & Grigorenko, E. L. (2000). *Practical intelligence in everyday life.* Cambridge: Cambridge University Press.

Sulsky, L. M. & Balzer, W. K. (1988). The meaning and measurement of performance rating accuracy: Some methodological concerns. *Journal of Applied Psychology, 73,* 501-510.

Sulsky, L. M. & Day, D. V. (1992). Frame-of-reference training and cognitive categorization: An empirical investigation of rater memory issues. *Journal of Applied Psychology. 77*(4), 501-510.

Tabachnick, B. G. & Fidell, L. S. (1983). *Using multivariate statistics*. New York: Harper & Row.

Task Force on Assessment Center Standards. (1989). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management, 18,* 457-470.

Taylor, P., Keelty, Y. & McDonnell, B. (2002). Evolving personnel selection practices in New Zealand organisations and recruitment firms. *New Zealand Journal of Psychology, 31*(1), 8-18.

Tedeschi, J. T. & Riess, M. (1981). Self-presentational styles. In J. T. Tedeschi (Ed.), *Impression management theory and social psychological research* (pp 3-20). New York: Academic Press.

Tenopyr, M. (1977). Content-construct confusion. *Personnel Psychology, 30,* 47-54.

Tett, R. P. & Guterman, H. A. (2000). Situation trait relevance, trait expression, and cross-situational consistency: Testing a principle of trait activation. *Journal of Research in Personality, 34,* 397-423.

Thornton, G. C. III (1992). *Assessment centers in human resource management.* New York: Addison-Wesley.

Thornton, G. C. III & Byham, W. C. (1982). *Assessment centers and managerial performance.* San Diego, CA: Academic Press.

Thornton, G. C. III, Kaman, V., Layer, S., & Larsh, S. (1995, May). *Effectiveness of two forms of assessment center feedback: Attribute feedback and task feedback.* Paper presented at the 23rd International Congress on the Assessment Center Method, Kansas City, Kansas.

Thornton, G. C. III, Tziner, A., Dahan, M., Clevenger, J. P. & Meir, E. (1997). Construct validity of assessment center judgments. *Journal of Social Behavior and Personality, 12,* 109-128.

Ting, N., Burdick, R. K., Graybill, F. A. Jeyaratnam, S. & Lu, T. C. (1990). Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computational Simulation, 35,* 135-143.

Tovey, R. C. (2001). *Anxiety and assessment centre performance.* Unpublished doctoral dissertation, Goldsmiths College, University of London, New Cross, London.

Turnage, J. J. & Muchinsky, P. M. (1982). Trans-situational variability in human performance with assessment centers. *Organizational Behavior and Human Performance, 30,* 174-200.

Turnage, J. J. & Muchinsky, P. M. (1984). A comparison of the predictive validity of assessment center evaluations versus traditional measures in forecasting supervisory job performance: Interpretive implications of criterion distortion for the assessment paradigm. *Journal of Applied Psychology, 69,* 595-602.

Wagner, R. K. (1985). *Tacit knowledge inventory for managers: Test booklet.* Unpublished manuscript, Department of Psychology, Florida State University, Tallahassee, Florida 32306-1270.

Wagner, R. K. & Sternberg, R. J. (1985). Practical intelligence in real-world pursuits: The role of tacit knowledge. *Journal of Personality and Social Psychology, 49,* 436-458.

Wagner, R. K. & Sternberg, R. J. (1986). Tacit knowledge and intelligence in the everyday world. In R. K. Wagner & R. J. Sternberg (Eds.), *Practical intelligence: Nature and origins of competence in the everyday world* (pp. 51-83). Cambridge: Cambridge University Press.

Wagner, R. K. & Sternberg, R. J. (1990). Street smarts. In K. E. Clark & M. B. Clark (Eds.), *Measures of Leadership* (pp. 493-504). West Orange, NJ: Leadership Library of America.

Wagner, R. K. & Sternberg, R. J. (1991). *Tacit knowledge inventory for managers.* San Antonio: Harcourt Brace & Company.

Warech, M. A., Smither, J. W., Reilly, R. R., Millsap, R. E. & Reilly, S. P. (1998). Self-monitoring and 360-degree ratings. *Leadership Quarterly, 9*(4), 449-473.

Whitmore, M. D. & Klimoski, R. J. (1984). *Leader emergence and self-monitoring behavior under conditions of high and low motivation.* Paper presented at the annual meetings of the Midwestern Psychological Association, Chicago.

Williams, K. M. & Crafts, J. L. (1997). Inductive job analysis: The job/task inventory method. In D. L. Whetzel & G. R. Wheaton (Eds.), *Applied measurement methods in industrial psychology* (pp 51-88). Palo Alto, CA: Davies-Black Publishing.

Woodruff, S. & Cashman, J. (1993). Task, domain, and general efficacy: A reexamination of the self-efficacy scale. *Psychological Reports, 72,* 423-432.

Woodruffe, C. (1993). *Assessment centres: Identifying and developing competence* (2nd ed.). London: Institute of Personnel Development.

Appendix I: Pilot For Study Three

Method

Participants

Data were collected from a development centre (DC) that was constructed for a large call centre in a government-based organisation in Auckland, New Zealand. The centre was used three times over a two-year period between 2001 and 2002 for the training and development of call centre workers. Fifteen organisational members participated, consisting of 11 females and four males with ages ranging from between 26 and 30. Nationality was not recorded, as it was felt the small numbers used in the present study could lead to the identification of individuals. All respondents reported that they held bursary (high school leaving) qualifications.

Assessors

The assessors were 5 managerial staff members per DC from a government-based organisation (1 male and 4 females), with a mean age of 31.60 (*SD* 5.24) located in Auckland, New Zealand. The assessors remained the same throughout the duration of the 3 runs of the DC, except for the last two runs in which 2 assessors had to be replaced. All assessors had previous experience in assessing participants in multiple ACs for selection, although none had previously received either FOR or psychological training. All participants had considerable experience (over 2 years), and were regarded as subject matter experts of the position being assessed.

The DC

After developing a policy statement outlining the purpose of this particular DC, and who would be involved in the process, the initial step in the construction of the DC was to execute a competency analysis of the target position. As the purpose of the current study was to compare a task-specific model with a dimension-specific model, the competency analysis involved a two-tiered process of producing a detailed task-analysis (gathering information on the tasks that organisational members perform), and then a classification and, moreover, an extrapolation of these tasks into dimensions. Inductive job analyses use various methods to find new and specific information about a given job (Peterson, & Jeanneret, 1997). This approach was taken in the present study for a number of reasons. Firstly, the intention of the study was to focus on the collection of new, detailed information about a particular job, in order to construct a unique and highly detailed account of the competencies involved in the job. Peterson and Jeanneret (1997) suggest that, in such situations, inductive methods are more appropriate, rather than the deductive methods which yield more general information. Additionally, as there were small numbers of subject matter experts (SMEs) in this sample, using job analysis questionnaires may have been problematic in terms of displaying high levels of error and inflated standard deviations reflected in job analysis questionnaires, which may have otherwise been abated with larger numbers. The last reason was that the particular organisation involved in this study had it's own agenda as to its required developmental specifications. Thus, as the will and developmental needs of the organisation was of great consequence in the construction of the DC, it was felt that such information should

be driven to some degree by the subject matter experts who had knowledge of the areas that required performance development.

Task Analysis

The first stage of the competency analysis involved utilising job analysis methods to identify the key tasks that made up the call centre position. This involved a review of the current job descriptions already in existence, interviews and critical incident interviews with incumbents and supervisors; the SMEs. The SMEs group (which comprised the same sample as the assessors) was interviewed. To give the analysis a strategic outlook, interviewees were also asked to give their views on what tasks they thought might be important for the call centre position in the future, as suggested by Thornton (1992) and Woodruffe (1993). In accordance with the guidelines set out by Lowry (1997), a questionnaire was developed listing the tasks derived from the information obtained in the task analysis to determine the relative rank and importance of particular tasks. Three questions were asked of SMEs with respect to each task, including: (a) the criticality of this task relative to others for successful operations, (b) time spent on this task relative to other tasks, and (c) the difficulty of this task relative to others. The last item differed from Lowry's suggested third item (relative importance of being able to perform this task correctly on entry into the job). This was because the intention of the present DC was for development, and it was reasoned, in agreement with the subject matter experts, that job entry requirements were not involved in developmental aims. Incumbents were already familiar with the task, and the importance

of tasks for entry level may be important for recruitment and selection, but this may not be relevant to development.

From this information, the most critical tasks were selected for inclusion in the DC exercises. In concurrence with the suggestions of SMEs, a checklist of the typical actions that would be required to successfully perform each of the tasks was developed. These typically involved short checklists of around 8-15 actions considered important for the successful completion of a given DC exercise. As Lowry emphasised, no inference of the existence of complex constructs was made at this stage, and it was at this point that the competency analysis for the task-specific DC model concluded.

Classification and Extrapolation of Tasks into Dimensions

In a traditional dimension-specific DC, the pure or raw tasks obtained from the task analysis are then classified into dimension categories. This involves a process of subjectively identifying and then coding the task with a dimension that is thought to underpin the performance of that task (Ballantyne & Povah, 1995). A dimension was assigned for all tasks with guidance from the generic dimensions suggested by Thornton and Byham (1982), and in concurrence with SMEs in the present study. Because some evidence suggests that using small numbers of performance dimensions may act to increase the construct validity of DCs (Lievens, 1998), the current DC followed the growing body of literature suggesting that DC architects should limit the number of dimensions assessed (Gaugler & Thornton, 1989; Lievens & Klimoski, 2001; Sackett & Hackel, 1979). Upon reviewing the literature, Arthur et al. (2000) decided on a manageable set of 9 performance dimensions, in line with human information processing

capacity. Note that Arthur et al. cited Miller (1956) on the issue of cognitive capacity. Arthur et al. looked at the number of dimensions assessed over 19 DC studies and found that on average, 11.01 (SD = 5.24) dimensions were assessed across the processes. The present DC, in the light of these findings, limited the number of performance dimensions to seven. These top seven dimensions were identified by the SME panel as being the most important for the purposes of the current development process. This number was within the optimal limits suggested by Gaugler and Thornton (1989) of between 5 to 7 dimensions.

DC Task Ratings and Dimensions

Task checklists were also developed with specific behavioural indicators of successful performance on an exercise were provided for assessors to mark. Assessors marked each specific task on a scale ranging from 1 (Performance was very much below standard) to 5 (Performance was very much above standard). Each task statement had a dimension name written next to it, to give the assessors guidance on which specific behaviours might relate to which dimension or competency trait.

Participants were rated on the following 7 dimensions: Process Utilisation; Conflict Resolution; Communication; Technical and Professional Knowledge; Customer Service Orientation; Stress Tolerance and Innovation. It was the intention to assess all dimensions across all exercises, except for Customer Service Orientation and Conflict Resolution which were not formally assessed in the Group Analysis Exercise. Spaces for marks for these dimensions were left on the forms for the raters. At the end of DC, it was evident that as a group, the raters felt that the Group Analysis Exercise afforded

opportunities for participants to manifest behavioural examples of Customer Service Orientation and Conflict Resolution. It was found that 80% of the raters had included ratings for Customer Service Orientation and Conflict Resolution for the Group Analysis Exercise.

These data were included in the analysis, and thus produced a fully crossed dimension by exercise design for analysis. Dimensions were assessed on a scale ranging from 1 (unacceptable level of ability) to 5 (very high level of ability). The following definitions were assigned to these dimensions:

Process Utilisation: The extent to which an individual gains as much benefit as possible from their use of existing resources.

Conflict Resolution: The extent to which a CSR can effectively manage a situation so as to diffuse the escalation of conflict.

Communication: The extent to which an individual effectively and accurately conveys oral or written information and responds to questions and challenges.

Technical and Professional Knowledge: The level of understanding of relevant technical and professional information.

Customer Service Orientation: The extent to which an individual is willing to provide proactive, efficient and effective fulfillment of customer requests over and above expectations.

Stress Tolerance: The extent to which an individual maintains a consistent level of performance under the stress of confrontation, tight time-frames and/or uncertainty.

Innovation: The extent to which an individual generates new or creative ideas and solutions, and uses available resources in new and more efficient ways.

DC Exercises

Four simulation exercises were employed to assess the tasks and dimensions in the DC. Three of these were high-fidelity call centre simulations that aimed to simulate calls from customers who had specific challenging issues that the CSR had to resolve. All exercises were set up so that the CSRs were positioned at computers which had standard databases installed, mirroring the computers that the CSRs had been trained on in their actual positions. Each computer was linked to a telephone station, where a role player sat. Each role player had been given a script, and was instructed to keep to the script as much as possible during the exercise. At an appointed time, the role payers called the assessment stations for each CSR. One assessor per CSR was assigned for the first three exercises. The last simulation was a lower fidelity group-exercise, where two assessors were assigned to one CSR. The simulations included the following exercises:

The Walkway Simulation: Portrayed a situation where a dissatisfied customer was calling about a large hedge that was blocking a walkway that the customer frequented. To add further challenge, the customer could not remember the specific name of the location of the walkway, nor were they aware of the actual definition of the term 'walkway'.

Recycling Bin Simulation: Involved another simulation where a dissatisfied customer gave the CSR unnecessary information, from which the CSR was expected to extract the information necessary to resolve the real issue that the customer had. The central issue involved the replacement of a government-owned recycling bin.

The Rates Simulation: Involved an inquiry into rates. Three specific issues needed to be contended with, including answering a customer enquiry relating to how rates were calculated and what rates actually paid for; payment options and changing addresses.

The Group Analysis Exercise: Attempted to assess an individual's contribution to a group exercise relating to a job relevant scenario. The scenario involved a customer email enquiry into rates, parks, rubbish collections, out-of-zone areas, and disaster information. The participant was rated on their input into discussions on the issues, utilisation of computer resources, and utilisation of the Internet to solve the issues presented.

Evaluation Approach

Several studies have sought to evaluate the relative efficacy of evaluating performance dimensions after the completion of each exercise (within-exercise rating), or waiting until the completion of the entire DC, and then making an evaluation of the dimensions concerned (within-dimension rating). As previously discussed, the evidence for the efficacy of one approach over the other remains unclear (Harris, et al., 1993; Silverman, et al., 1986). As the two approaches appear to contribute relatively little to the facilitation of the construct validation of the DC process, it was decided that the within-exercise approach would be used. Additionally, this approach was used because the DC used in this study was developmental in nature. Feedback, therefore, needed to be given to participants as an ongoing process throughout the DC. As previously discussed, there is a large body of evidence to suggest that DCs typically show more method than dimensional variance in ratings. To ensure that participants were given appropriate feedback, the within-exercise approach was favoured. Participants were

given feedback on behaviours (tasks pertaining to a particular exercise) as suggested by Lowry (1997). Also, feedback was given on the basis of ability traits that were assessed in particular exercises, rather than giving feedback to candidates on the basis of dimensions assessed across the different exercises, as suggested by Feltham (1989). If ability dimensions in DCs were to be conceptualised as relatively stable, enduring characteristics, then theoretically, they should stand up to being rated in individual exercises (as in Campbell & Fiske, 1959). Note that in any case, the treatment of ratings with respect to feedback, and for determining OARs was secondary to the principle aims of the study.

Assessor Training and the Assessment Procedure

Assessors were trained on the DC exercises using behavioural observation training (Ballantyne & Povah, 1995) coupled with guidance on how to use behavioural checklists to assist the process (Lowry, 1997). It has been suggested that frame of reference (FOR) training (Bernardin & Buckley, 1981) enhances the appraisal of human performance (Murphy & Cleveland, 1995). Additionally FOR training has been suggested as a factor that may act to increase the construct and criterion validity of DC ratings (Arthur, et al.; Lievens, 1998; Schleicher, 1999). The present study used frame of reference notions as an integral aspect of the training procedure.

Prior to the DC, assessors were trained in how to assess participants using a frame of reference training procedure that has been suggested by Lievens (1998) for use with DCs. This involved a training session with assessors that covered some basic principles in assessing behaviour, and familiarised the assessors with the exercises and the rating

instruments that would be used. The FOR component of the training involved having assessors rate the performance of assesses on a contrived CSR written about in a short vignette. Both behavioural and trait ratings were then displayed on a white board together with the mean and standard deviation of the ratings provided by each assessor. The assessors were then invited to discuss the ratings they had given. These discussions focussed on relatively large standard deviations, and why some raters might deviate from others, in the hope that a shared schema could be constructed for what was construed as good versus poor performance on a given exercise. This procedure was an abbreviated version of procedures that have been recommended in the literature (Lievens, 2001a), due to the strict time demands enforced by the organisation under study. The suggested FOR format was followed more closely in Study Three proper.

In sequence, the process involved firstly a general explanation of the DC process, and the benefits to the organisation of utilising this procedure. Next, a general description was given of the two that would be involved in the assessment process, including the observation and rating of behaviours, and from those behaviours, the inference of dimensions or traits could theoretically be made. In congruence with the guidelines of Ballantyne and Povah (1995), assessors were then shown how to assess behaviours with no construct inference. This process involved observation, and the recording of behaviours on notepaper and then a checklist (Lowry, 1997) to obtain a score relating to behavioural performance. This increased objectivity, and guided the scoring process, whilst allowing for a numerical rating to be allocated to the key behaviours in the DC. The inference of ability traits over and above these behavioural

ratings was the next stage of classification. Here, assessors were shown examples of the behaviours representing, or theoretically underlying, each dimension.

For each exercise, assessors were trained to rate behaviours first using the behavioural rating checklist. Each behaviour was denoted as being a possible underlying factor for a superordinate dimension. Assessors gave an inferred score for these dimensions on a 5-point scale (Ballantyne & Povah, 1995) within each exercise. Assessors were then trained in the consensus discussion procedure for dimensional ratings, where at the conclusion of the DC, all assessors presented evidence and critically discussed the ratings they had obtained to form OARs for each participant. Assessors looked at each participant individually, and assessed their performance on each dimension individually. Evaluation on each dimension was backed up by reported behavioural observations by each assigned assessor for each exercise and each participant.

Once the procedure was completed, assessors gained mastery experiences (Bandura, 1982; 1986) through rating role players in two simulated DC exercises: the walkway exercise, and the recycling bin exercise. This allowed an opportunity for assessors to compile their own behavioural ratings, from which they extrapolated trait ratings. As Arthur et al. suggested, the assessors then discussed their findings to work towards building a common frame of reference for performance on the exercise or dimension.

Procedure

The DC ran between late 2001 and early 2002 for the period of about ½ a day for three separate sessions. The centre was run according to a schedule where each group of participants performed each exercise in turn. Each candidate was given behavioural and trait-based ratings on their performance during the DC. At the conclusion of each exercise, participants were given coaching feedback by their assessors on their performance, and what they could have done to improve their performance on a given simulation exercise.

Results and Discussion

Although the government-based organisation in the present study was originally intended as a full investigation into DC ratings, the organisation under scrutiny opted out of the project after the DC had been constructed, assessor training had been completed, and 15 participants had completed the DC. It was decided that while the results of 15 participants could not possibly constitute a meaningful investigation into DC ratings, the sample could act as a pilot study, and indeed a great deal of information, in terms of process improvement, was gained from this precursor. The reader is urged not to draw conclusions based on the following analyses. The results of this pilot should be regarded as a learning device and a precursor to the actual Study Three. A briefer version of the analysis presented in Study Three is, therefore, presented in this pilot study.

The data for the Pilot to Study Three were imputed for missing data using EM (expectation maximisation), which uses an iterative process, by which to estimate missing values. Out of a total of 1380 potential scores across the two DCs, eight behavioural

scores and one trait score were missing. Of an additional set of two traits that were included for analysis, eight scores were missing. More detail on this addition is given below. Thus, in total, 15 scores were missing (nearly a 99% response rate).

As previously discussed in the method section, it should be noted that some data were added to the total set, which were not originally intended for inclusion. The two traits 'conflict resolution' and 'customer service orientation' were not originally intended for assessment in the group discussion exercise in the DC. The subject matter experts who rated the DC argued that they saw manifestations of these traits in the exercise, and 80% of the raters scored these traits anyway. It was decided that these ratings should be included, as DC design commonly relies heavily on the opinions of subject matter experts (Ballantyne & Povah, 1995; Lowry, 1997) and the situation was beneficial for the ANOVA used in a G study. With the inclusion of these additional ratings, exercises and traits could be considered fully crossed, which meant that the variance attributed to exercises and traits could be considered independently.

Table 49 shows the grand means and standard deviations for the task-specific DC presented for each exercise. Under the task-specific approach, performance on particular exercises is considered the most important unit of measurement. All mean scores vacillated around the 2nd and 3rd points on the rating scale. Standard deviations for the task-specific ratings fluctuated around one rating. Table 50 shows the grand means and standard deviations for the dimension specific DC. Under the dimension-specific approach, performance on particular dimensions is considered the most important unit of measurement. Like the task-specific DC, average dimensional

Table 49

Grand Means and SDs of the Behavioural Ratings (Within Exercises) in the Task-Specific DC

Exercise	*M*	*SD*
Walkway	2.10	1.00
Recycling Bin	2.50	0.97
Rates	2.45	1.01
Group Exercise	2.12	0.97

ratings centred around the 2nd and 3rd points on the rating scale. The mean for the last dimension 'Innovation' was slightly lower than the others at 1.88. Standard deviations for the dimension-specific ratings fluctuated around one rating.

The present study employed Generalizability Theory (G theory, see Brennan, 2001a; Cronbach, Gleser, Nanda & Rajaratnam, 1972) to analyse data. G studies utilise variance components models that are derived from the mean squares calculated in factorial ANOVAs. Although statistical significance is not generally considered to be of importance in G theory (Brennan, 2000), the confidence limits within which one computes estimates of components of variance can be calculated using confidence intervals designed specifically for variance component estimates (Brennan, 2001a). Such confidence intervals cannot be theoretically justified for designs that are unbalanced with respect to nesting (Brennan, 2001b). The task-specific DC was unbalanced with respect to nesting to the effect that it was not viable to extract items

Table 50

Means and SDs of the Dimension ratings (Across Exercises) in the Dimension Specific DC

Dimension	*M*	*SD*
Process Utilisation	2.50	0.70
Conflict Resolution	2.53	0.79
Communication	2.62	0.92
Technical and Professional Knowledge	2.67	0.77
Customer Service Orientation	2.72	1.41
Stress Tolerance	2.45	0.81
Innovation	1.88	0.87

in order to contrive a balanced design. Therefore, confidence intervals were not calculated for the task-specific design. Confidence intervals were, however, calculated for the fully-crossed dimension specific design.

The effects of differences between raters both within and between the DCs were not thought to be of great concern in the present study, because the same raters were used for the same participants across the two DCs. This was so that the effect attributable to raters would be held constant over the task-specific and dimension-specific administrations. Also, different raters assessed different participants in a rotation system in accordance with the suggestions of Lievens (1998). It was hoped that such a system would randomise error associated with rater idiosyncrasy to the greatest extent possible. However, during the course of the DC, this allocation was not always systematic as raters changed their order and some assessors rated more

participants than others. The complexities of the unsystematic nesting of assessors disallowed their inclusion in the present G study. To gain an estimate of interrater reliability, equation 1,1 from Shrout and Fleiss (1979) was employed for each DC. Equation 1,1 was relevant to the present sample because each participant was rated by a random combination of assessors who were selected from a larger population of judges.

Specific facets were included in the G study that were instrumental in addressing the research issue at hand. The task-specific DC was a partially nested design, in that each exercise had its own specific set of items. The facets included in the task-specific process included exercises (*e*), items nested within exercises (*i*:*e*), and an estimate of the variance attributable to the object of measurement, persons (*p*). All interaction terms were also analysed. The dimension-specific DC employed a fully crossed design incorporating the facets exercises (*e*), dimensions (*d*), and an estimate of the variance attributable to the object of measurement, persons (*p*). All interaction terms were also analysed.

Table 51 shows the G study for the fully balanced comparison between the task-specific and dimension-specific DCs. All variance components and confidence intervals were calculated using urGenova ver. 2.1 (Brennan, 2001b). Listed for each type of DC are the object of measurement, facets, and interactions (effects), degrees of freedom (*df*), variance component estimates (VC), 90% confidence intervals and the percent of explained variance (explained variance %) as a heuristic for identifying the proportional contribution of various facets to variation in scores (Shavelson & Webb, 1991). While the effects in the task-specific DC *x*, *i*:*x* and in the dimension specific DC *x*, *d*, and *xd* are presented in Table 51, these facets alone provide little

Table 51

Pilot Generalizability Study Comparing a Task-Specific with a Dimension-Specific DC in a Repeated Measures Design

	Task-Specific DC					Dimension-Specific DC			
Effect	*df*	VC	90% Confidence Intervals	Explained Variance (%)	Effect	*df*	VC	90% Confidence Intervals	Explained Variance (%)
p(persons)	14	0.0202	*	2.0	*p*(persons)	14	0.0774	0.0008 < VC < 0.2570	10.2
x(exercises)	3	0.0139		1.4	*x*(exercises)	3	0.0207	0.0000 < VC < 0.3521	2.7
i(items):*x*	60	0.1879		18.4	*d*(dimensions)	6	0.0694	0.0276 < VC < 0.2778	9.1
px	42	0.2682		26.3	*px*	42	0.2801	0.1904 < VC < 0.4386	36.8
pi:*x*,*e*	840	0.5295		51.9	*pd*	84	0.0068	0.0000 < VC < 0.0337	0.9
					xd	18	0.0144	0.0011 < VC < 0.0455	1.9
					pxd,e	252	0.2928	0.2544 < VC < 0.3412	38.4

Note: Confidence intervals were calculated using the Ting et al. (1990) procedure described in Brennan (2001a). Ting et al's procedure is recommended for random, balanced designs so as to avoid the computation of inaccurately wide intervals. * Confidence intervals were not provided for the task-specific procedure because the specification of a confidence interval for an unbalanced design is inappropriate (Brennan, 2001b). The task-specific design in this case was too unbalanced to viably contrive a balanced design by removing items.

information of interest to the present study. As with any form of assessment in the selection context, the focus is on person variation across the various facets, because of the notion that assessment procedures aim to discriminate between people for decision purposes. Therefore, the focus in the present study concerns interactions between persons and facets and variance component estimates for the object of measurement. Of particular interest in the present study is the interaction term *px* for both types of DC (Kane, 1982; Kraiger & Teachout, 1990; Lievens, 2001b). In the task-specific approach, the *px* interaction was a comparatively high contributor, explaining 26.3% of the variance in the model. A proportionately high *px* interaction in the task-specific approach is defined by variation in the candidate's performance according to different situations (exercises) presented to them. In the dimension-specific approach, the *px* interaction was also comparatively high at 36.8%. Again, this interaction reflects the extent to which candidate performance varied across exercises.

The interaction term *pd*, in the dimension-specific DC, reflects the extent to which dimensions are useful for discriminating between persons (Lievens, 2001a; 2001b). This interaction term explained 1.0% of the variance in the dimension specific model. Additionally, the effect for the object of measurement, *p*, was estimated for the task-specific approach, and explained 2.0% of the total variance. The object of measurement, *p*, for the dimension specific approach was higher, and explained 10.2% of the variance in scores. This was thought to be influenced by training and design issues that needed rectification. Also, the lack of person discriminability in the task-specific approach could have been influenced by poor interrater reliability discussed later. Indeed, person variation was poorly estimated in the dimension-specific study, as evidenced by the corresponding wide confidence interval in Table 51. The terms *pi:x,e* and *pxd,e* in the task-specific and dimension-

specific processes, respectively, are difficult to interpret purely as they contain the interactions between all facets and the object of measurement in the model, plus undifferentiated random error.

Confidence intervals are presented in Table 51 for the dimension-specific model. All confidence intervals were calculated using the method suggested by Ting, Burdick, Graybill, Jeyaratnam, and, Lu (1990), which is generally recommended for random, balanced designs so as to avoid the computation of inaccurately wide intervals for variance component estimates (Brennan, 2001a). Table 51 suggests that particular variance component estimates in the dimension-specific model were poorly estimated as evidenced by wide confidence intervals, including the effect for *p*, *x*, and *px* in particular. In all probability, poor estimation was also theoretically obtained for the task-specific approach. These are further reasons that the reader should not place a great deal of confidence in the findings from this study.

G theory acknowledges that in practice, relative and absolute decisions are often made about individuals on the basis of a psychological measure. A relative decision is one in which the performance of individuals are compared with other individuals (e.g., norm comparisons present relative decisions where people are compared with one another). An absolute decision is one in which a certain cut-off criterion is employed (e.g., a pass or fail criterion for employment decisions). G theory provides two coefficients for the purposes of relative and absolute decisions that are analogous to reliability coefficients in classical test theory. Tables 54 and 55 provide the equations and calculations, for both types of DC, of σ^2_{Rel} (relative error; all of the effects in the G study that contribute variance to relative decisions), σ^2_{Abs} (absolute error; all of the effects in the G study that contribute variance to absolute decisions), $E\rho^2_{Rel}$ (the Generalizability or G coefficient; for relative decisions), and ϕ

(Phi coefficient; for absolute decisions). Tables 54 and 55 also provide equation ICC 1,1 from Shrout and Fleiss (1979) as an estimate of interrater reliability across the two types of DC.

Table 52 shows that for relative decisions, $E\rho^2_{Rel}$ was calculated at 0.23, and for absolute decisions, ϕ was calculated at 0.21 for the task-specific model. Additionally, ICC 1,1 was calculated as 0.42 for the task-specific model. For the dimension-specific model in Table 53, $E\rho^2_{Rel}$ was calculated at 0.49, ϕ was calculated at 0.44, and ICC 1,1 was calculated as 0.45. It should be noted by the reader that

Table 52

Relative and Absolute Error, Generalizability and Phi Coefficients and Interrater Reliability for the Pilot Task-Specific DC

Index	*Result*
$\sigma^2_{Rel} = \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_{pi:x,e}}{n_{i:x}n_x}$	0.07
$\sigma^2_{Abs} = \frac{\sigma^2_{x}}{n_x} + \frac{\sigma^2_{i:x}}{n_{i:x}} + \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_{pi:x,e}}{n_{i:x}n_x}$	0.08
$E\rho^2_{Rel} = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{Rel})}$	0.23
$\phi = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{Abs})}$	0.21
$ICC(1,1) = \frac{BMS - WMS}{BMS + (k-1)WMS.}$	0.42

Table 53

Relative and Absolute Error, Generalizability and Phi Coefficients and Interrater Reliability for the Pilot Dimension-Specific DC

Index	*Result*
$\sigma^2_{\text{Rel}} = \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_d}{n_d} + \frac{\sigma^2_{pxd,e}}{n_x n_d}$	0.08
$\sigma^2_{\text{Abs}} = \frac{\sigma^2_x}{n_x} + \frac{\sigma^2_d}{n_d} + \frac{\sigma^2_{px}}{n_x} + \frac{\sigma^2_{pd}}{n_d} + \frac{\sigma^2_{xd}}{n_x n_d} + \frac{\sigma^2_{pxd,e}}{n_x n_d}$	0.10
$E\rho^2_{\text{Rel}} = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{\text{Rel}})}$	0.49
$\phi = \frac{\sigma^2_p}{(\sigma^2_p + \sigma^2_{\text{Abs}})}$	0.44
$\text{ICC}(1,1) = \frac{\text{BMS} - \text{WMS}}{\text{BMS} + (k-1)\text{WMS.}}$	0.45

given the results of the G study, the use of $E\rho^2_{\text{Rel}}$ and ϕ in this context is somewhat debateable. It is argued in the original monograph on G theory "While it is not assumed that *p* [the variance attributable to the object of measurement] is completely stable during the period to which the universe definition applies, it is taken for granted that *p*'s characteristics fluctuate around a typical value" (Cronbach et al., 1972, p. 363). That is to say, there is at least some stability of responding assumed when employing G and Phi. The use of these coefficients is perhaps questionable because the evidence from the G study suggests, in line with previous research, that the DCs ratings reflect situationally specific responses, rather than stable characteristics.

However, Cronbach et al. suggest that when the occasions of assessment are considered as samples of behaviour, it is "mathematically sound to define the universe score as the average over the time span [over which behavioural measurements will be made]" (p. 363). This might reflect overall performance on the exercises as samples of behavioural performance, a conception that seems acceptable in the role of task-specific DCs, where it is necessary to pool results at the end of the process to provide a summary rating for selection purposes (Lowry, 1997).

Again, the low results for the indices presented in Tables 54 and 55 suggest that the dependability and reliability of measurement in the pilot study was low, and therefore should not be used for decision-making purposes. The pilot study did; however, lead to process gains, and aided the researcher in developing the AC in Study Three.

Appendix II: Introduction to Generalizability Theory

Study Three utilised Generalizability Theory (G theory) (Cronbach, Gleser, Nanda & Rajaratnam, 1972) as a paradigm under which to analyse assessment centre data. This paper provides an opportunity to elucidate G theory for those not accustomed to its alternative view on the concept of dependability and reliability. The focus in this short paper is not to provide a comprehensive account of what has become the holistic tapestry that G theory is today. Such an account is, to date, most fully described in Brennan (2001) and Marcoulides (1998). Rather, the focus is on some of the theoretical aspects of G theory that are often not dealt with in-depth, and to aid the reader to form a conceptual grounding that will aid an interpretation and understanding of the foundations of G theory.

G Theory

The Theoretical Stance Underlying G Theory

Cronbach et al. (1972) originally conceptualized G theory as a model for understanding the dependability of behavioral measurements. They remarked, "The decision maker is almost never interested in the response given to the particular stimulus, objects or questions, to the particular tester, at the particular moment of testing. Some, at least, of these conditions of measurement could be altered without making the score any less acceptable to the decision maker" (p. 15). Thus, it is the *score* that is considered integral in G theory. The means by which the individual came to earn that score are considered exchangeable with some other, just as acceptable, means. To illustrate:

consider an item on a given test. G theory suggests that this item might just as easily be replaced with any other item that could reasonably be expected to measure the same construct. The test designer would deem such an alternative item acceptable. Thus, Cronbach et al. maintain "The ideal datum on which to base the decision would be something like the person's mean score over all acceptable observations" (p. 15).

Under the notions presented above, G theory presents an alternative view of the dependability of psychological measurement. A dependable measure, under this viewpoint, is one that can accurately generalize from a person's observed score on a test, to that person's mean score under all possible conditions that would be acceptable to the test user or decision maker. The interest lies in obtaining a dependable score for a person here: the means by which the person came to gain that score can be altered and changed. In this sense, the question asked by G theory is 'Can this person's score, that is, the observed score, generalize to an idealistic score that reflects that person's average over all the possible conditions under which this score could be obtained?' The idealistic score mentioned here is a hypothetical construct, called a universe score.

Note that a person's measured attributes are considered relatively stable and enduring under this paradigm, i.e., as though they were trait-based, and differences in scores across different occasions of measurement, e.g., across items in a test, or across exercises in an AC, are attributable to one or several sources of error. Both items and exercises from the previous example could be considered as potential sources of error variance. It is these sources of error that G theory first attempts to isolate, and then quantify in terms of their relative contribution to the variance in the scores gained by a person.

Trait-based vs. Situationally Specific Forms Of Assessment

Because G theory assumes some kind of situational stability in responding, measures that are intended for responses to specific situations, e.g., task-specific ACs (Lowry, 1997) or work sample exercises, become theoretically problematic on first inspection. Such forms of assessment are task based, in that they do not make the inference of any stable underlying characteristics inherent within an individual, and are often employed in the practice of personnel psychology (Schmidt & Hunter, 1998). As will be seen later, this possible limitation is not problematic when one is at the stage of identifying the various sources of error that contribute to scores. That is to say, regardless of any trait-based assumptions, G studies can be performed on practically any personnel data.

The only time when G theory becomes conceptually challenging, in this regard, is when generalizability coefficients are calculated. It is argued in the original monograph on G theory "While it is not assumed that *p* [the variance attributable to the object of measurement] is completely stable during the period to which the universe definition applies, it is taken for granted that *p*'s characteristics fluctuate around a typical value" (p. 363). That is to say, there is at least some stability of responding assumed when employing G theory.

This could be regarded as a limitation of the G study approach when it comes to analyzing task-specific ratings, in that the expected score, in G theory, under any condition is assumed, to some degree, to be the same. Consider the previous example of the task-specific AC, in which assessment exercises are treated as though they are stand-

alone work samples of situationally specific behavior. Cronbach et al. suggest that when the occasions of assessment are considered as samples of behavior, it is "mathematically sound to define the universe score as the average over the time span [over which behavioral measurements will be made]" (p. 363). This might reflect overall performance on the exercises as samples of behavioral performance, a conception that seems acceptable in the role of task-specific ACs, where it is necessary to pool results at the end of the process to provide a summary rating for selection purposes (Lowry, 1997).

As Cronbach et al. mention, the concept of a universe score becomes dubious when an individual's performance is changing appreciably in a regular trend. Certainly no regular trend, for instance performance worsening or improving dramatically, is necessarily intended in a task-specific AC. The wider intention of G theory is to identify relatively stable differences between people on the basis of some measure. Because task-specific ACs include an overall score, it could be argued that there is some general level assumed in performance across exercises that contain similar assessment content. This does not infer the existence of a trait; indeed, it is not necessary to make such an inference in behavioral model under which task-specific forms of assessment operate. Rather, this could be conceptualized as a general response to a set of readily exchangeable situations that contend with similar subject matter. In effect, the logic presented here suggests that even with situationally based responding, similar situations will tend to elicit responses from individuals that could be seen to fluctuate around a typical value.

The fact remains that the subject matter across the exercises in a task-specific AC (i.e., the situations) are likely to hold similarities. Ahmed, Payne, and Whiddett (1997)

suggest in their guidelines for AC exercise construction that the exercises should be related to one another. As such, there might be some generality in responses to such similar situations, i.e., a universe of similar responses elicited by similar situations exist for the type of circumstances assessed. Thus, under a behavioral paradigm, it is arguable that a person's behavior will fluctuate to some degree around a central value, in a task-specific AC, if the situations hold similar characteristics. Indeed the task-specific ACs in the present dissertation hold very similar characteristics across exercises. Thus, in keeping with the assumptions of Cronbach and his colleagues, G theory should be applicable even to task-specific ACs of this type.

As an aside, given the assumptions of G theory, one might ask why multiple exercises are included in an AC, when one exercise might suffice. This argument goes back to a paradox in classical test theory that is made clear through G theory. In classical test theory, one could quite possibly increase the reliability of an AC by reducing the number of exercises, even down to a singular exercise. The less variance attributable to different exercises in this model, the higher the reliability of measurement. This would, in all probability, lead test designers to feel insecure with the assessment of an individual, because the assessment would be confined to the idiosyncrasies of a particular exercise. In G theory, the concept of reliability resolves into an argument for the accuracy of generalization. One exercise will generalize accurately to a very narrow universe: a universe pertaining to a certain type of exercise. The use of multiple exercises will allow generalization to a much more important universe in practice: a universe of the use of multiple exercises for assessment (Shavelson, Webb & Rowley, 1989).

G Studies

G studies utilize factorial ANOVA models to derive a comprehensive dissemination of the facets that contribute to variance in the scores obtained on a measure.

Factorial ANOVA

A fundamental tool in univariate G theory is factorial ANOVA. Factorial ANOVA can be used to partition the variance in scores into various components. The variables that contribute to variance are called 'factors' in ANOVA and 'facets' in G theory. G theory uses the term 'facet' as opposed to 'factor' to avoid evoking associations with factor analysis (Cronbach et al., 1972). The variance components that are calculated can be used to indicate the relative contribution of a particular facet, or the interactions between multiple facets, to scores. Factorial ANOVA looks at the variance components attributable to singular facets (main effects) and interactions, as well as all of the facets in the specified model in combination with one another. Some of these constructs can be isolated as contributors to error variance.

The term that is identified for the interaction between all of the facets in a model is usually defined as the error term, and represents the effect for all of the interactions, plus undifferentiated error. Undifferentiated error is defined by contributors to variance that are unsystematic and are unable to be isolated. For example, someone might be distracted during their completion of a personality test by a loud noise. The loud noise thus presents an uncontrolled source of error variance that is unsystematic and therefore

cannot be accounted for. Factorial ANOVA is used as the tool with which G studies separate potential systematic sources of variance. G studies use the information from a factorial ANOVA to partition error variances and to calculate coefficients, including G Coefficients.

Facets in Generalizability Theory

In contrast to classical test theory, which is confined to estimating true scores and then combines together all sources of error variance, G theory aims to isolate individual contributors (facets) to the error variance in scores in a single analysis. Indeed, it is this simultaneous partitioning of the sources of error variance that distinguishes G theory from Classical Test Theory. The individual sources of error variance found in a G study can then be used to glean information about how to maximize the dependability of a particular test or measure in a Decision study (D study), by calculating various Generalizability Coefficients. Aspects of D studies are discussed later.

The Universe of Admissible Observations

G theory defines what is labeled a *universe of admissible observations.* The universe of admissible observations is a set of "observations that a decision maker is willing to treat as interchangeable for the purposes of making a decision" (Shavelson & Webb, 1991, p. 3). Thus, it is the wider set of observations that a test user would find equally acceptable for a given purpose (Cronbach et. al, 1972). Any given observation is treated as a sample from the theoretical universe of observations deemed admissible by a test user or test developer. Note that G theory specifically uses the term *population* to

describe a set of subjects or participants, and uses the term *universe* to describe a set of facets (Cronbach et. al, 1972).

To exemplify, consider a universe that has one facet, an identified source of error, called items. The universe of admissible observations in this case would be the potentially endless set of items that could replace the observed set of items, i.e., the items currently in the test, with the caveat that it must be reasonable to assume that all of these items measure the same construct. That is, they would need to be deemed admissible by the test developer, or test user. Other generalizations about facets can be made in similar ways. There might be a universe of possible forms of a test, or a universe of potential test administrators. For instance, if a test measures intelligence, the score attributed to the internal attribute "intelligence" is thought not to be restricted to the results of one test. It is presumed that the aspects of the test should generalize to aspects of tests purporting to measure the same construct. If this generalization is made, then there is evidence that the test is dependable or generalizable (hence the term 'Generalizability Theory').

It is the facets of a test, (e.g., items, forms, administrators) which can lead to errors in generalizing from the test to the universe. Take 'items' for example. If all of the items in the universe of admissible observations for items tend to measure the same trait, and a person's score on those trait items are similar, then one might expect generalization from a sample of those items to a universe of those items. If the items are not measuring the same construct, and a person's scores differ enormously from one item to the next, generalization from the sample to the universe will be hazardous. This will lead to error in generalizations made about an individual's level on a particular measure. Thus, if items do generalize from a sample of test items to a universe of items deemed to

be measuring the same construct, then assumptions can be made as to the efficacy of a test in terms of its ability to make generalizations about a person's level of a particular construct.

As a concrete example, consider an AC. When conducting a G study on an AC, one would specify the universe of admissible observations broadly, so as to encompass as many facets as possible. This is so that the chosen model reflects the reality of the measurement device and so that one can identify which facets actually contributed to the variance in scores. Note that the broader the definition, the more sources of variance that are included in the assessment practice, the more difficult it will be to generalize from the sample to G theory's theoretical ideal score, the universe score variance.

The universe of admissible observations for a given study, whatever its definition, must reflect the set of observations that would be equally acceptable for the test user's purpose. It is an operational definition of the class of procedures considered in the measurement model (Cronbach et al, 1972). A less elegant way of describing this term would be to label it *the perpetual set of exchangeable conditions of facets* which implies in the same way that there is a larger set of conditions of facets that could theoretically be exchanged with the ones actually observed. Defining the universe of admissible observations is all about specifying which facets should be included in a study. A universe of admissible observations can be defined by one facet, two facets, or more. The more facets included in the model, the more complex the model becomes.

Firstly, a researcher might reason that the different traits specified for measurement in this particular AC might produce error variance in scores. Thus, 'traits' can be specified as the first facet. Secondly, it could be argued that different raters might

produce error variance in scores. Therefore, 'raters' could convincingly become the second facet. Error variance might also be attributable to different simulation exercises that are used in an AC. 'Exercises' would be the third facet. Similarly, the different occasions on which an AC is run might present some form of error variance. 'Occasions' becomes the fourth facet.

The facets prescribed or specified in a G study define the universe of admissible observations. Thus, the model described above presents a complex model to prescribe for a G study. The definition of the universe of admissible observations in this AC would be defined by all acceptable traits that could be assessed by all acceptable raters across all acceptable exercises at all acceptable points in time. As can be seen, this definition could easily apply to nearly any dimension-specific AC. There could also be other facets that might sensibly be included in the model.

G studies not only consider the main effects of all of the facets incorporated into a model, but also look at all the possible interactions that could occur between them. As an example, one might consider the effect of an interaction between different raters and different occasions. Interaction effects reveal that main effects are modified by the presence of interactions with other facets in the specified model. They suggest that the main effect cannot be interpreted alone, but should be considered also in terms of its relationship with other facets. It might be that one AC was run on Thursday, and another was run on Friday. On Friday, the raters as a group did not concentrate properly due to eager feelings with regard to the potential activities of the coming weekend. Thus, it is likely that in this case, there will be an interaction between the effects of raters and occasions because rater behavior altered across different days.

The Object of Measurement

Another important aspect that has not yet been considered is that pertaining to the effect of the *object of measurement*. Indeed the effect obtained for the variance attributed to the object of measurement is an integral component in G theory. The object of measurement is the person, animal, or object that is actually being observed and rated. In studies of I/O psychology, this is usually the variance component or effect attributable to persons, or participants. The object of measurement in a G study is initially treated in the same way as the facets are treated: as a source of variance. G studies also look at the interaction between the object of measurement and the other sources of variance in the model. Fundamentally, under the G theory paradigm, the variance attributed to the main effect of the object of measurement is not considered as a source of measurement error.

The whole aim, intention and meaning behind the study of individual differences is to evaluate diversity across individuals on the basis of certain measured characteristics. Psychological tests and ACs constitute popular methods by which to assess individual differences in I/O psychology. Thus, the variance arising from differences between the objects of measurement will define a crucial element of G theory. This will be detailed in the section dealing with G Coefficients.

Crossed and Nested Designs for G Studies

Two kinds of research designs are generally considered by G theory; *crossed and nested designs*. A crossed design occurs when every condition of one facet is observed with every condition of another facet. For example if, in the AC mentioned earlier, every trait were assessed in every exercise, this would mean that traits and exercises were

crossed. This is because every condition of one facet (traits) was observed with every condition of another facet (exercises). Crossed designs are more desirable because they ensure that the individual effects of the facets can be separated from one another. From the above example, one would be able to differentiate the individual influence that traits and exercises had on the ratings in the AC.

The second type of design, a nested design, occurs when two or more conditions of one facet occur with only one specific condition of another facet. To illustrate, if it was decided that three ACs would be run over the course of three days, different participants could be evaluated on each day that the AC was run. Thus, each day will have its own specific set of participants. In such a scenario, participants are said to be nested within days.

In effect, nesting produces independent groups that could each contribute to variation in scores. Nested designs are less desirable than crossed designs in G theory, because if one facet is nested within another, it becomes difficult to disentangle the individual effects of the nested facet. The effect of the nested facet becomes inextricably linked with the facet within which it is nested. As such, one cannot obtain a clear idea of the individual influence of the nested facet. However, nested designs are often chosen out of practicality. Crossed designs are often by no means practical, however they yield a richer analysis. There is a trade-off when choosing either form of research design.

In the specification of designs for analysis, crossed and nested facets utilize certain symbols to indicate their status. When a facet is crossed with another, the symbolization for persons crossed with test items (i.e., every person completed every test item) would look like: $p \times i$ (in that p = persons and i = items). If items were nested

within people (i.e., particular groups of people completed particular groups of items) the symbolization would look like: *i:p*. If *p* and *i* were the only facets to be included in the model, then the error term would look like *pi,e* where *e* indicates undifferentiated error. Each facet has a variance component attached to it in a G study. For the calculation of variance components, the interested reader is directed to Shavelson and Webb (1991). The SPSS or SAS Windows based statistical programs can also compute variance components for G studies. GENOVA, a DOS based program devoted to research using G theory, is also available for these calculations.

Random and Fixed Facets Under G Theory

G theory takes a distinctive perspective on what it considers to be a random and a fixed sample. It is important to note that in G theory generally, most facets are assumed to have been sampled at random, and thus G theory is, essentially, a random effects model. If a facet is considered to have been sampled at random, then the sample is smaller than the universe of that facet. Take, for example, an AC that has three different simulation exercises. G theory will ordinarily treat these exercises as though they have been sampled from a possibly endless universe of simulation exercises that could have potentially been used in the AC. Thus, exercises would ordinarily be considered as a random facet, contingent on the nature of the exercises, and the extent to which the justification for defining them as random is cogent.

Shavelson, Webb, and Rowley (1989) warn that any inference that is made from the sample should be only directed at the population from which that sample was drawn. An argument that is often employed to justify G theory's assumption of random variables

comes from Bayes' Theorem, from a notion labeled *exchangeability*. This concept suggests that although the facets have not been sampled in a purely random manner, they may be considered as being sampled at random if the facets that are *not* included in a given G study could be exchanged with or are equally acceptable in comparison to the facets that *are* included in the G study (Shavelson & Webb, 1981; Shavelson, Webb & Rowley, 1989; Shavelson & Webb, 1991). Thus, if the designer of an AC would be content with exchanging the exercises in the AC with some other exercises that might perform the same function (at the same level of acceptability), the exercise facet can be considered as being sampled at random. This is an assumption that is made by the theory from the outset, and could present a possible limitation in the theory. One would not realize the true reality of the nature of the exchangeability of the facets without further research into this notion.

A facet is considered fixed in G theory when the conditions relating to it exhaust all of the conditions in the universe of generalization. Thus generalization from the sample to the universe is not relevant because the entire universe has already been captured by the conditions of the facet. For example, consider research on the effect of the day of the week on AC ratings. If every day of the week were included in the facet, days of the week would need to be considered as a fixed variable because there would not be any other conditions (i.e., days) to make generalizations to.

D Studies

D studies utilize the information gleaned from G studies, to make decisions about the dependability of a given measure. While the purpose of a G study is to estimate

variance components; the purpose of a D study is to estimate quantities specific to a particular measurement procedure, and its relationship to a universe of generalization.

Relative and Absolute Decisions

D studies use two different kinds of coefficient that pertain to two different kinds of decision that a test user may wish to engage in. Both of these decisions have wide applications in employment. The first is referred to as a *relative decision.* This involves situations where the decision maker is interested in how the individual performed relative to other people. This is analogous to the concept of using norms, where one might claim that an individual scored higher than 60% of his or her peer group.

The second kind of decision that D studies acknowledge is that pertaining to *absolute decisions.* In G theory, absolute decisions are ones in which no comparison to any peer group is necessitated. These are decisions where a person either passes or fails, or is awarded some score on the basis of a criterion that has nothing to do with the relative standing of individuals. An example of this might be a driving test, where the criterion is set for a person to pass if they answer more than 90% of the test items correctly. This score has nothing to do with how others have performed on the test. It is an absolute decision as opposed to a relative one. Brennan and Kane (1977) are credited with some aspects of applying G theory to absolute decisions.

Universes of Generalization

One of the most important considerations in a D study concerns the universe to which a researcher wishes to generalize, on the basis of the results derived from a

particular measure (Brennan, 2001). The universe of generalization is defined as the specific universe to which the researcher wishes to make generalizations to. This consideration relates whether a given measurement model is considered random or fixed. In sum, considerations given to the universe of generalization inquire as to whether the researcher wishes to generalize to a much larger group. To illustrate, for a development center, the researcher might be interested in generalizing from the scores obtained on the basis of exercises and dimensions used in the process, to those same scores obtained on a greater population of exercises, and dimensions. This model, as mentioned earlier, is considered random. The universe of generalization will be the direct consideration when calculating and interpreting G Coefficients.

The Generalizability Coefficient

Closely related to the notion of the universe of generalization is the *Generalizability Coefficient* (G Coefficient). On a 0-1 scale, a G Coefficient reflects the likelihood that the measure will be able to locate individuals relative to other members in the population. Thus, the G Coefficient focuses on the object of measurement, which usually constitutes individuals. The G Coefficient represents how generalizable the score for an individual would be over exhaustive measurement in a measurement model. Universe score variance is the variance attributable to the ideal score that one wishes to obtain. This ideal score is the average score that an individual would obtain across all the possible measurement conditions in the universe of admissible observations in a particular measurement model.

In the original monograph written on G theory, Cronbach and his colleagues (1972) stated "the tester is interested chiefly in the person tested and only secondarily in the conditions of observation" (p. 2). As stated earlier, the variance in scores that is attributable to the object of measurement, usually the person tested, is fundamental to G theory. In fact, G theory uses the variance attributable to the object of measurement to estimate universe score variance when calculating a G Coefficient. The variance component for the object of measurement is considered as a representative sample of that object of measurement in the universe. This is considered as the numerator in the calculation of the G Coefficient.

As previously mentioned, it could be argued that it is desirable to explain variance through certain facets, commonly through trait-based dimensions in an AC. However, the ultimate aim in the study of individual difference is to locate disparity between individuals in order to characterize their various areas of strength and weakness. The means by which the tester came to conclusions about the differences between individuals are considered secondary to the point that disparity was actually found. The denominator in the G Coefficient reflects those secondary sources of variance. This is labeled 'expected observed-score variance', and is estimated by combining the variance component for the object of measurement with the other sources of measurement variance included in the definition of the universe of admissible observations. The choice of effects included in the denominator of this equation depends on the type of decision to be made.

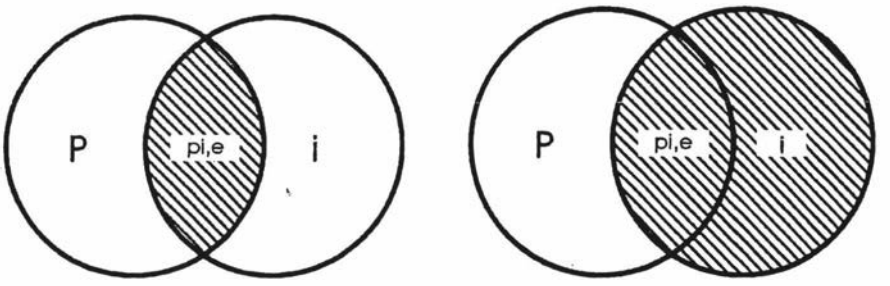
There are two kinds of G Coefficient, the choice of which depends on the type of decision that will be made with a particular assessment process. As mentioned

previously, G theory recognizes two such decisions: relative and absolute. Error variance is different for the two kinds of decision, and therefore the G Coefficient is calculated differently for one decision over another.

Measurement Error

To calculate a G Coefficient, two indices of measurement error are initially calculated for inclusion into the generalizability coefficient formulae, for each respective decision. The facets contributing to variance for relative decisions include all of the interactions between the object of measurement and the facets, plus undifferentiated error. This does not include the variance component for the object of measurement, which as mentioned previously, is an estimate of universe score variance. Relative error includes all of the interactions showing how people differed with each other on the various facets. These features will affect the relative standing of individuals.

For a random model and absolute decisions, all of the variance components in the model except the variance component for the object of measurement are included in the reliability formula. Figure 1 shows sources of error for relative and absolute decisions in a random design where persons (*p*) are crossed with items (*i*), taken from an example in Shavelson and Webb (1991, p. 86). The facet *pi,e* refers to the interaction between persons and items, together with undifferentiated error (*e*). The shaded parts indicate which components should be included in the calculation of measurement error for each respective decision. The concepts presented in Figure 1 can be taken as a rule of thumb, and although the design in Figure 1 is reasonably simplistic, the rules are applicable to other, more complex designs.



Relative Error **Absolute Error**

Figure 1. Sources of error for Relative and Absolute Decisions for a Random *p* X *i* Design.
Note: From Generalizability Theory (p. 86) by R. J. Shavelson & N. M. Webb, 1991, CA: Sage Publications. Copyright 1991, Sage Publications. Reprinted with permission.

For a relative decision (σ^2Rel) with the same design as in Figure 1, the equation for the estimated relative error variance would be:

$$\sigma^2_{\text{Rel}} = \frac{\sigma^2 pi,e}{n'_i} \qquad [1]$$

For an absolute decision (σ^2Abs) again with the same design as in Figure 1, the equation for the estimated absolute error variance would be:

$$\sigma^2_{\text{Abs}} = \frac{\sigma^2_i}{n'_i} + \frac{\sigma^2_{pi,e}}{n'_i} \qquad [2]$$

The formulas described here can apply to any, more complex universe of admissible observations. Note that the symbol n' is a G theory symbol by convention, which, in this case, means the number of items that will be included in a D study. The calculations of either absolute or relative error (or both), depending on the decision to be made with the assessment data, will be used to calculate the G Coefficient.

Calculating a Generalizability Coefficient

For relative decisions, a G Coefficient is calculated by dividing the variance component that was obtained for the object of measurement (construed as universe score variance) by the variance component attributable the object of measurement plus the measurement error calculated for relative decisions. For absolute decisions, a G Coefficient, or more correctly a Phi Coefficient, is calculated by dividing the variance component that was obtained for the object of measurement, construed as universe score variance, by the variance component attributable the object of measurement plus the measurement error calculated for absolute decisions. Note, as an aside, that a G Coefficient, using relative error variance, is an example of a Pearson-developed intraclass correlation.

For a relative decision, the formula for a G Coefficient for any universe of admissible observations is defined by the following:

$$E\rho^2_{Rel} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{Rel}} \qquad [3]$$

For absolute decisions for any universe of admissible observations, the formula for a G Coefficient is defined by the following:

$$\phi = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{Abs}} \qquad [4]$$

One aspect of a G Coefficient is reminiscent of the Spearman-Brown prophecy formula, which allows the prediction of reliability on the basis of test length, that is, the number of items in a test. D studies, however, go over and above the Spearman-Brown

formula because not only can they use the number of items as a criterion to maximize the reliability of a given measure, but they can also use all of the other identified sources of variance as contributors to this prediction as well. To contextualise, G theory forms the theoretical basis for G studies, the results of which can be used to calculate G Coefficients in a D study. D studies utilize comprehensive information about a psychological measure to find means of improving such a measure. In a D study, G Coefficients are calculated. These coefficients form the basis for the decisions made about psychological measures in D studies.

Some Considerations For The Application Of G Theory

Data obtained for use in G theory should be at interval level, or at least ordinal in nature. Variance components are subject to sampling variability, thus, one must be careful about making generalizations on the basis of the sample used. Factorial ANOVAs do not assume any particular distributional form of data since statistical significance is often not calculated. Significance is a concept that is often not considered relevant in G theory. Brennan (2000) argued that the absence of any assumption of the distribution of data is a strength of G theory, as normality based assumptions associated with other procedures are highly suspect. However, when maximum restricted likelihood procedures are used (Brennan, 1992) and when confidence intervals relating to variance components are calculated (see Brennan, 2001), the assumption of a normal distribution is indeed made. Consideration also needs to be given to whether a design is balanced or not (see Brennan, 2001).

The applicability of D studies is bound to some degree by the nature of the design of the original D study. Thus, crossed designs are far more flexible to the G theory researcher, although they are not always workable in practice. Take for example a fully crossed AC process, notably with raters crossed with participants crossed with exercises. This would involve having every participant rated by every rater across every assessment exercise. Clearly this situation would be impractical.

Some concerns have been raised with regard to the calculation of negative variance components when using factorial ANOVA models. Indeed, factorial ANOVAs are sensitive to issues such as the number of subjects associated with levels of a particular facet and small sample sizes in general (Shavelson, et al., 1989). These factors, in combination with misspecification of a particular measurement model, can influence the calculation of redundant or negative variance components. In such cases where a negative variance component is found, attention should be drawn to issues regarding sample size and the correctness of the specification of the factorial model. With respect to the latter, factorial models can be complex, and require a great deal of forethought with regard to the measurement process that was followed.

For situations where the negative variance component sustains despite the above considerations, the negative component may arise from sampling error. Cronbach et al. (1972) suggest setting the negative component to 0, and then using 0 in any following calculations. Brennan (1992) also suggests setting the negative variance component to zero, however, the suggestion is made that the negative estimate should be used in all calculations relating to that component. The first approach is potentially biased, whilst it casts aside a notion that is conceptually problematic. The second approach utilizes a

conceptually problematic notion whilst minimizing bias. One of the alternative methods of calculating variance components, which does not return negative variance components, is restricted maximum likelihood estimation (Shavelson & Webb, 1991).

Other assumptions inherent in factorial ANOVA include independence of the effects specified in the factorial model, and the assumption that the same standard error of measure is often applied to all objects of measurement. As Strube (2000) comments, the latter of these assumptions is unlikely to be true. Consideration should also be given to the arguments of Schmidt and Hunter (1996). They suggest that flawed conclusions may be drawn from studies that do not consider important sources of error. Consideration should be given to potentially 'hidden' facets. To illustrate, consider one large AC that was administered at one occasion. 'Occasion' in this study is a constant, and would not be considered as a source of variance in the study. However, in practice, it is feasible that smaller ACs might run over the course of several occasions. Such a situation presents what is known as *transient error*, and reflects that a given G study has not accounted for a potentially important source of variance, that could be influential when making decisions on the basis of a given measurement procedure.

Situations may arise when more blatant or important sources of error are omitted out of reasons of practicality, or because the necessary information was not available. For example, in the AC examples in the present dissertation, the source of error attributable to assessors was not included. This was because assessors were not systematically allocated (as detailed in the results section of Study Three). Depending on the variability across raters in an assessment situation 'assessors' may constitute an important source of variance that has not been acknowledged. Strube (2000) suggests

that fundamentally, the calculation of universe score variance and error variance depend on both conditions of measurement and conditions of application. Thus, the universe of generalization is very much dependent on sources of error that are included in a G study.

The limitations of G theory are likely outweighed by its potential benefits. G theory is far less restrictive and much more comprehensive than its classical counterpart. Work to improve G theory is constantly underway (Brennan, 2001). With greater understanding and application, G theory may become the force guiding measurement in a variety of employment scenarios. The aim is to increase the precision and confidence with which decisions are made on the basis of different modes of psychological measure.

Arguments in favor of the use of G Theory for AC Data and General Testing In Employment.

G theory has been suggested by Lievens (1998) as an appropriate method for use when attempting to understand the sources of variance that contribute to AC ratings. Lievens argued in favor of G theory because of the holistic nature of the approach (e.g., in terms of explaining the variance in scores that might be attributable to particular components in a measurement model). In keeping with the position of Arthur, Woehr, and Maldegen (2000) and Lievens (1998), it is argued that in contrast to other methods of construct related assessment, G theory allows for a clear separation of the sources of variance underlying the multitrait-multimethod data in ACs.

As noted by previous authors (Arthur, et. al, 2000; Turnage & Muchinsky, 1982) a clear separation of the factors that underlie AC scores has been lacking in many research publications. Factor analysis is a commonly used approach, however, this

technique tends to focus primarily on the influence of exercises and traits, and neglects other potentially important sources of variation and the interactions between these sources of variance. G theory, however, does allow for the simultaneous differentiation of many potential sources of variance that might contribute to scores in an AC. For example, the influence of different occasions at which an AC took place and their contribution to score variance, or the interaction between different occasions and the different exercises. These general arguments also apply to general tests that are developed for employment. G theory can be utilized to maximize measurement and decision accuracy by the researcher (Lievens, 2001). Such accuracy is vital in employment contexts where researchers and practitioners are making decisions that will affect the course of people's lives.

References

Ahmed, Y., Payne, T. & Whiddett, S. (1997). A process for assessment exercise design: A model of best practice. *International Journal of Selection and Assessment, 5*(1), 62-68.

Arthur, W., Woehr, D. J. & Maldegen, R. (2000). Convergent and discriminant validity of AC dimensions: A conceptual and empirical re-examination of the AC construct-related validity paradox. *Journal of Management, 26*(4), 813-835.

Brennan, R. L. (1992). *Elements of generalizability theory* (rev. ed.). Iowa City, IA: American College Testing Program.

Brennan, R. L. (2000). (Mis)Conceptions about generalizability theory. *Educational Measurement: Issues and Practice, 19*(1), 5-10.

Brennan, R. L. (2001). *Generalizability Theory.* New York: Springer-Verlag.

Brennan, R. L. & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277-289.

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: John Wiley.

Lievens, F. (1998). Factors which improve the construct validity of ACs: A review. International Journal of Selection and Assessment, 6, 141-152.

Lievens, F. (2001). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology, 86*(2), 255-264.

Lowry, P. E. (1997). The assessment center process: New directions. In R.E. Riggio & B.T. Mayes (Eds.), Assessment centers: Research and applications [Special issue]. Journal of Social Behavior and Personality, 12(5), 53-62.

Marcoulides, G. A. (1998). Applied generalizability theory models. In G. A. Marcoulides (Ed.). *Modern methods for business research* (pp. 1-21). New Jersey: Lawrence Erlbaum.

Schmidt, F. L. & Hunter, J. E. (1996). Measurement error in psychological research: Lessons from 26 research scenarios. *Psychological Methods, 1,* 199-233.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124,* 262-274.

Shavelson, R. J. & Webb, N. M. (1981). Generalizability theory: 1973-1980. *British Journal of Mathematical and Statistical Psychology, 34,* 133-166.

Shavelson, R. J. & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

Shavelson, R. J., Webb, N. M. & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44*(6), 922-932.

Shavelson, R. J. & Webb, N. M. (1991). Generalizability theory: A primer. Newbury Park, CA: Sage Publications.

Strube, M. (2000). Reliability and generalizability theory. In L.G. Grimm & P.R. Yarnold (Eds.). *Reading and understanding more multivariate statistics* (pp. 23-66). Washington, DC: American Psychological Association.

Turnage, J. J. & Muchinsky, P. M. (1982). Trans-situational variability in human performance with ACs. *Organizational Behavior and Human* Performance, 30, 174-200.

Appendix III: Abridged Assessment Centre Manual and Training Guide For Farmers Merchandiser, General Sales and One On One Sales Roles: Including General Sales Exercises

Abridged Assessment Centre Manual and Training Guide For Farmers Merchandiser, General Sales and One On One Sales Roles: Including General Sales Exercises

Duncan J. R. Jackson

Assessor Training

Preparing oneself for the use and administration of the present assessment centre will initially entail reading over the material presented in this document (the declarative component). The next section (the procedural component) will utilise a practical behavioural observation and frame of reference training program. The behavioural observation component will form the foundation of the assessment process. This will begin with training in observing behaviours, recording behaviours and using behavioural checklists to assist when classifying behaviours. In the behavioural checklists, there are recommendations as to which behaviour (or action) might relate to which underlying competency. These recommendations are provided in parentheses after each action detailed in the behavioural checklists. The classification of behaviours into competencies will largely be a judgmental process, based on the evidence the assessor has gained from the behavioural ratings.

The underlying notion or assumption is that seen behaviours (factual observations) are manifestations of an unseen underlying psychological competency (or ability trait). For example, if we see someone speak clearly across many different situations, we might make the inference that the person holds a relatively stable and enduring characteristic that we could label "Communication Ability". These ideas are shown in Figure 2.

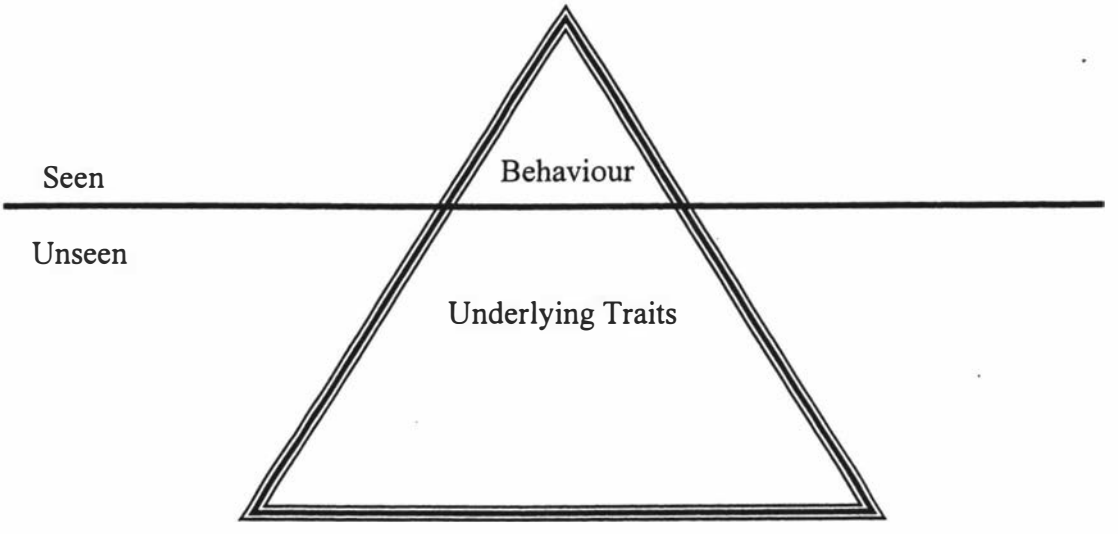


Figure 2: Iceberg Model of Behaviour (Based on Honey, 1986 & Fleenor, 1996)

Assessors will be provided with participant and assessor instructions, behavioural observation and behavioural checklists as well as a set of competency definitions. Prior to the assessment centre, assessors need to familiarise themselves with the tasks involved on the behavioural observation forms and the competency definitions. It is vital to have this information readily available in memory. Reading over these documents to familiarise yourself with them prior to the centre will assist your ability to assess enormously. The behavioural checklists include all of the expected actions the participant should execute in a given exercise. The competency definitions will give varying levels of the extent to which a participant possesses a particular competency.

a) Observing Behaviours

Assessors will need to observe and record behaviours. These processes are used simultaneously; behaviour is observed and then systematically, the assessor records **only** the behaviour they saw without making any inferences beyond that behaviour. In this sense, it is only the behaviour that we are interested in recording, and these pure observations should not be contaminated with judgmental comments, biases or inferences over and above the behaviour that is observed.

b) Errors / Considerations In Rating

Assessors need to be wary of several observational errors that may occur:

- The halo effect occurs when an assessor was overly influenced by a participant's performance on one part of the exercise. This could result in an assessor ascribing a particularly positive or negative all round account of a person, and basing this judgement on one characteristic or behaviour. It is vital to consider the entire range of behaviours when assessing, to avoid this potential source of error.
- Be aware of stereotyping in terms of prejudice towards specific individuals, and do not allow such subjective biases to influence what should be an objective rating procedure.
- Make sure that your observations on one exercise do not influence your observations on another. Treat the exercises independently of one another.
- Be aware of what a participant does not say and do, as this can provide potentially valuable information (e.g., what data were not used in solving a problem for a customer).
- Leniency/Harshness errors occur when assessors rate with unusual clemency or unusual severity, respectively.
- Central Tendency errors occur when assessors avoid extreme ratings, and tend to rate an individual with multiple middle scores. Ideally, an assessor should try to use the full range of ratings available.
- Recency and Primacy errors occur when most emphasis is put on either the most recent or the first behaviours seen (respectively). Be sure to take the whole spectrum of behaviours into account to avoid this.

c) Recording Behaviours

As previously mentioned, observation and recording of behaviour occur simultaneously, even though they are presented here as though they are two discrete stages. As a result, recording involves only the transcription of the events that occurred, with no interpretation of the meaning behind these events. In this sense, at both the observation and recording levels, the assessor acts as an objective data collector. The Interpretation of these data comes later.

The assessor records only the behaviour they saw without making any inferences beyond that behaviour. In this sense, it is only the behaviour that we are interested in recording, and these pure observations should not be contaminated with judgmental comments, biases or inferences over and above the behaviour that is observed. With behavioural observation, the concern is one of fact, not interpretation. For example, imagine a group discussion where one of the group members remains completely silent. It is easy for an observer to ascribe assumptions or inferences to this group member such as "she wasn't interested" or "he was obviously bored" or "she is introverted". These assumptions are based on pure inference, and at the behavioural observation level, inference needs to be avoided. Ballantyne and Povah (1995) liken the assessor's role as an observer as analogous to a video camera. The camera will only reflect what actually happened without interpretation. As an observer, it is necessary to record only what was observed, not what we thought the observation meant. It is necessary to take highly accurate and detailed notes of the events that transpired. If the observing and recording phases are impaired, then all the later stages will certainly be impaired also. There are some major points to remember when recording behaviour:

- Record only observable characteristics – what was actually seen and heard, do not make inferences over and above this information.
- When recording, write the time at regular intervals during the assessment.
- Note down all the actions that the candidate makes on the group exercises.
- Record as much non-verbal behaviour as possible (e.g., looking away, leaning forward).
- When dealing with verbal data, develop a form of short hand. In most cases, it will probably not be possible to note down all that was said or all the actions taken, but when it is not possible to write down such information verbatim, it is absolutely acceptable to write down key words in a sentence or key words pertaining to actions to give you the context of what was being said (e.g., But we didn't.....).
- Make sure what you observe and record is correct, as this process forms the basis for the entire assessment operation. The behavioural notes that have been taken will serve as the foundation information for classifying behaviour into Behavioural Checklists, and then into Competencies.
- Note: Make sure that as an assessor, you do not converse (other than introductions) with the participants, either before or during the assessment centre. Eliciting conversations could potentially lead to the development of unfair bias, or the divulgence of sensitive information.
- During the assessment centre, sustain a consistently professional manner (e.g., avoid joking) and do not converse with the participants.
- Ideally, no form of encouragement should be given. Standardisation is an issue

here: If one group receives encouragement and another does not, the assessment will be unfair.

- In situations where it is absolutely necessary (e.g., a quiet group), encouragement must ONLY be given by the administrator. All other assessors MUST remain silent.
- If discussions draw to a close prior to time, and reasonable discussion has taken place, it is acceptable to move on to the next part of the assessment.
- Note that one assessor is generally called upon to be the administrator. The administrator must read from the administration card on the assessment exercises **verbatim**. Again standardisation is at issue here. Avoid adding pieces to the administration, as this will make the process unfair and unreliable.
- The administrator must be firm with the participants, as well as very clear in their instructions.

d) Classifying Behaviours into Behavioural Checklists

After the behavioural notes have been taken, the next step in the process is to classify these into behavioural checklists. Classification is by nature, in an assessment centre, a judgmental process. However, be careful to base these judgements only on the objective data obtained from the behavioural notes. Initially, the behavioural notes that were taken at the recording phase need to be classified into the behavioural checklists provided for each exercise. It is useful at this time to highlight the important behaviours that were observed on the behavioural notes. Classification should ideally occur straight after the participant has completed an exercise, whilst the information is still fresh in the assessor's mind. At this stage of the process, behavioural notes and any written material provided by the candidate can be used for evidence in classification. The task here is to determine the relationship with the behaviours listed in the behavioural notes, with those presented in the behavioural checklist.

Mark a score for each expected action in the assessment centre on the behavioural checklist and the overall task performance in the exercise. *The overall task score need not reflect the average, but rather your overall judgement of the participant's task performance, again using the 6-point scale shown in figure 3.* Do not share this information with the other assessors, as each assessor should judge behaviour independently. Be thorough and careful when marking and analysing the important behaviours observed.

Several important points need to be given consideration when completing the behavioural observation checklists:

- Rate contextually. As the assessors hold intimate knowledge of the job, rate relative to all the people you have encountered in similar positions. Do not compare the participants with each other, but rather, with your own global standard. The practical training day will help you to realise what this standard actually is.
- At times, behaviour may be conflicting (i.e., effective at one time, non-effective at another time during one exercise). Score performance in accordance with the dominant behaviour and report conflicting behaviours.

- At times, participants may not demonstrate particular behaviours. In these cases, simply leave the relevant item on the behavioural checklist blank. Assessors are encouraged to concentrate hard, however, to find behavioural evidence.
- When this process is completed, mark an overall score on the basis of the behaviours observed based on your judgement, this need not reflect the average.
- All of the rating scales used in this assessment centre contain 6-points, and no fractional values are permitted. This is to encourage assessors to make a decision with regard to the candidates. As decisions must be made in terms of candidates applying for positions, middle level grades provide very little information on which to base a decision.

Use the following 6-point scale on the behavioural checklist for each action. **Do not use fractional values**. Performance ranges from being:

1. Certainly below standard
2. Somewhat below standard
3. Unsure, probably below standard
4. Unsure, probably above standard
5. Somewhat above standard
6. Certainly above standard

Figure 3: 6-point scale for behavioural and competency assessment

- **Step 1: Complete the behavioural checklist below, using your behavioural notes and any other material that has been completed by the participant.**

Table 1: Checklist, General Sales, Closing Simulation

Use the following 6-point scale on the behavioural checklist for each action. **Do not use fractional values**. Performance ranges from being:

1.	Certainly below standard
2.	Somewhat below standard
3.	Unsure, probably below standard
4.	Unsure, probably above standard
5.	Somewhat above standard
6.	Certainly above standard

	Score	Expected Action
1.	☐	Individually suggests closes that would be appropriate to the scenarios. (Customer Focus).
2.	☐	Assists the team by suggesting appropriate closes in a group situation (Teamwork, Customer Focus).
3.	☐	Identifies customer needs (Customer Focus).
4.	☐	Focuses on finding solutions that will assist the other team-members (Teamwork, Tolerance).
5.	☐	Interacts in a positive and polite manner with the other participants (Teamwork, Tolerance).
6.	☐	Acknowledges and encourages other members of the group (Teamwork).
7.	☐	Speaks clearly and annunciates appropriately (Oral Expression).
8.	☐	Appears to be content with the interactions in the group (i.e., the participant did not appear to become angry or frustrated in any way) (Teamwork, Tolerance).
9.	☐	Writes clear and concise notes during the exercise (Comprehension).
10.	☐	Follows instructions that are given to him/her (Comprehension).
11.	☐	Keeps a constant level of interpersonal effectiveness during the exercise (Tolerance).
	☐	From your judgement (i.e., not based on the average score) assign an *overall score* based on your perception of how well the candidate performed the actions required to complete this exercise on the behavioural 6-point scale. Do not use fractional values.

e) Classifying Behaviours into Competencies/Dimensions

After completing the behavioural checklists, the next step is to classify the tasks/behaviours into competencies. Again, this process relies on a degree of judgement, however, be certain that these judgements are based on sound and objective data, and are not contaminated with human bias. Classification in terms of competencies, or the underlying psychological drivers involved in behaviour, involves several important considerations:

- Make absolutely certain that prior to the practical training, and especially well in advance of the assessment centre, that you are familiar with the competency dimensions (i.e., what defines high and low performance on a particular competency). This information is given in the competency definitions.
- From your behavioural notes and behavioural checklists, identify the first piece of behaviour that you observed. This could be in the form of a short sentence, single word, or long paragraph.
- Upon identification of this behaviour, make an attempt to relate it to the competencies listed in the competency definitions. Make sure you are familiar with the definitions of the competencies so that you are able to relate a piece of behaviour to a given competency. It is common for several competencies to relate to one piece of behaviour.
- Also note that the omission of behaviour can constitute important evidence. For example, the participant may have been asked a direct question to which they gave no answer. The role of the assessor is to interpret what this behaviour meant in the context of the exercise, and in relation to the competencies listed.
- Decide whether the behaviour is a positive or negative example of the competency, and note this in your behavioural notes (see Figure 4 for an example).
- Also, use the behavioural checklists as evidence for your judgement. On the behavioural checklists, you will notice the competencies that are associated with each behaviour, stated in parentheses.
- The next step involves evaluating a competency. Look at all the positive and negative evidence identified in the behavioural notes and behavioural checklists, and assign a rating in accordance with the dimensions outlined in the competency profile. This is a judgmental process, but should form its basis from objective evidence.
- Repeat this process for each behaviour as you work through the transcript of your notes and your behavioural checklists. You will inevitably find some behaviours that are saliently more relevant than others.

Time	Participant 1	Participant 2
10am	From what I am reading here, it appears that the issue relates to colour schemes (C+ OE+)	
	I think that it would be best to ask the customer about what colour scheme they have in their house (CF+ OE+)	
		Hey that's a really good idea (TW+ OE+)
	Yes, and that way they could match up their colours, and they would actually get what they want (CF+ OE+)	
		That would keep everyone happy (TW+ OE+).
10:10am	Can't you come up with your own ideas? Why do you have to use mine all the time? (TW- T-)	
		Excuse me?? (OE+)

Where 1) C = Comprehension
2) TW = Teamwork
3) T = Tolerance
4) OE = Oral Expression
5) CF = Customer Focus
6) + = A positive example of a competency
7) - = A negative example of a competency

Figure 4: Example of Behavioural Notes with Competency Annotations.

Based on the competency definitions and the behavioural checklist for this exercise, Use the following 6-point scale for each competency. **Do not use fractional values**. Performance ranges from being:

1. Certainly below standard
2. Somewhat below standard
3. Unsure, probably below standard
4. Unsure, probably above standard
5. Somewhat above standard
6. Certainly above standard

Figure 5: 6-point Rating Scale for Competencies. Note this is the same as the scale for behaviours.

- **Step 2: Complete the competency profile rating form below using your behavioural notes and the behavioural checklist.**

Table 2: Competency Profile Rating Form

1. Closing Simulation

Based on the competency definitions and the behavioural checklist for this exercise, Use the following 6-point scale for each competency. **Do not use fractional values.** Performance ranges from being:

1. Certainly below standard
2. Somewhat below standard
3. Unsure, probably below standard
4. Unsure, probably above standard
5. Somewhat above standard
6. Certainly above standard

Competency	Rating
Comprehension	
Oral Expression	
Tolerance	
Teamwork	
Customer Focus	

Notes

Overall competency rating for this exercise based on your judgement (not the average), on the 6-point scale. Do not use fractional values.

Table 3: General Sales Participant Performance Matrix

PARTICIPANT:				
Overall Competency Ratings	**Exercises**			**Average Ratings**
	Approach	*Closing*	*Returns*	
Teamwork				
Customer Focus				
Oral Expression				
Tolerance				
Comprehension				

Overall Assessment Rating: ☐

PRESENTATION: Did the individual maintain a high level of personal presentation throughout the course of the assessment centre? (Please tick ONE)

- Yes ☐
- No ☐

SUITABILITY: Based on the evidence gained in the assessment centre, is the individual suitable and well matched for the position they were assessed for? (Please tick ONE)

- Suitable ☐
- Middling ☐
- Unsuitable ☐

f) Summary

Here is an over-simplified step-by-step summary of what you will be doing in the assessment centre.

1. The assessment centre will begin. Two participants will be assigned to you, so make sure you know who they are and make sure you are in a position where you can see their responses.
2. On standard notepaper, you will take notes on the responses of the individuals you are assessing.
3. Immediately after each exercise, you will be allocated time to complete the ratings scales for the individuals you assessed. For ***each individual***, you will use the notes you took in step 2 to:
    a. Complete the behavioural checklist for the particular exercise (see Table 1). Then using the evidence obtained from the notes and the behavioural checklist, you will need to:
    b. Complete the competency profile rating form (see Table 2).
4. At the very end of the assessment centre, the overall ratings for each competency will be transferred from the competency profile rating form (see Table 2) to the participant performance matrix for the particular role (see Table 3).

The standards of what defines these different levels of behaviour will be dealt with in further detail in the later section on Frame of Reference Training. In general, the aim is to assess the participant in terms of their relative performance compared to other individuals in the position. In this sense, the ratings given are contextual.

g) Integration of Ratings

Commonly, assessment centres use assessor integration discussions to obtain overall scores with regard to an individual. Overall scores represent how well a person performed, on the whole, in the assessment centre. Due to practical considerations, the present assessment centres will utilise the calculation of average ratings across behaviours and competencies. Such processes have been validated in the literature (Pynes and Bernardin, 1992) and are accepted in the international guidelines and ethical considerations for assessment centres (International Task Force on Assessment Center Guidelines, 2000). It would be far too time consuming, in the present assessment process, to discuss the number of candidates that participate in the assessment centre individually.

Practice Assessment

To familiarise assessors with the assessment process it is important that they receive some practical experience with the assessment tools. This section will comprise most of what is covered in the training course.

- **Frame of Reference Training Procedure**

The training course will cover frame of reference training: a procedure that has been found to consistently increase the accuracy of assessor judgements by facilitating the development of a common frame of reference or shared mental model as to what constitutes high and low performance when assessing assessment centre participants. Assessors will be presented with some practice groups to assess. All assessors will ascribe both behavioural *and* competency ratings to *one* participant in the practice runs of the assessment centre. Three exercises will be assessed per assessment centre, and in each exercise, a different participant will be assessed. Ratings will be compared to ensure that all assessors share a similar mental model with regard to the varying levels of performance. It is important to be familiar with the behavioural checklists for each exercise and the competency definitions prior to training.

Following will be presented all of the rating scales that you will need to complete the frame of reference training procedure. The instructor will guide you on what you will need to do.

APPENDIX

Each competency is rated in accordance with the following scale. Specific behavioural examples of the competencies are included in behavioural checklists in the assessment centre exercises.

Use the following 6-point scale on the behavioural checklist for each competency. **Do not use fractional values**. Performance ranges from being:

1.	Certainly below standard
2.	Somewhat below standard
3.	Unsure, probably below standard
4.	Unsure, probably above standard
5.	Somewhat above standard
6.	Certainly above standard

Teamwork:
The extent to which the individual works effectively and harmoniously with other team members.

Customer Focus:
The extent to which the individual attempts to assist customers to make satisfactory purchases, is concerned with customer needs, describes products accurately, and matches presentations to the customer's interests.

Oral Expression:
The extent to which the individual speaks grammatically and clearly in appropriate language and using appropriate gestures.

Tolerance:
The extent to which the individual interacts effectively with people despite delicate, frustrating or tense situations that demand understanding, patience and empathy.

Comprehension:
The extent to which the individual understands spoken and written, verbal, or behavioural language.

Farmers Merchandiser Competency Definitions

Each competency is rated in accordance with the following scale. Specific behavioural examples of the competencies are included in behavioural checklists in the assessment centre exercises.

Use the following 6-point scale on the behavioural checklist for each competency. **Do not use fractional values**. Performance ranges from being:

1. Certainly below standard
2. Somewhat below standard
3. Unsure, probably below standard
4. Unsure, probably above standard
5. Somewhat above standard
6. Certainly above standard

Teamwork:
The extent to which the individual works effectively and harmoniously with other team members.

Problem Solving:
The extent to which the individual applies reasoning, thinking, and analytical ability.

Oral Expression:
The extent to which the individual speaks grammatically and clearly in appropriate language and using appropriate gestures.

Adaptability: Repetition:
The extent to which the individual adjusts to repetitive and simple activities without becoming dissatisfied or losing efficiency.

Vision:
The extent to which the individual observes differences between details in colour, form or visual appeal, ranging from large and rough guesses to small and exact features.

Note to the instructor: Attach Relevant Exercises Here

Approach Exercise

(A Group Analysis Exercise)

The Approach exercise aims to elicit key characteristics that are considered important for both the One On One Salesperson role and the General Sales role. The exercise also views individual behaviour whilst interacting in a group environment.

Scenario

The Approach exercise presents 3 situations where a typified customer enters a store. The group of participants are to devise a plan as to the best method of approach that should be applied to the individual customers.

Say to the participants:

"This is an exercise that presents several different types of customers that you might expect to find when you are working in a department store. Your task is to imagine you are a team of salespeople. You will need to listen carefully to each of the scenarios that I will read to you. Then, for each scenario I present, you will have 2 minutes to write some notes on how you think the customer should be approached. You will need to take notes as you will be assessed on them. You will then have 5 minutes to reach a consensus with the other group members as to what you think will be the one best way of approaching the customers I read about."

Then ask:

"Is everyone clear on what will happen?"

[Deal with any questions appropriately]

"Please listen carefully as I read the first scenario to you.

Scenario One

A very well dressed customer walks into the store and moves towards the computer area. Computers are your specialty. Your first impression is that the customer looks as if they want to make a purchase.

"For 2 minutes, take some brief notes on how you think this customer should be approached. Start now."

When 2 minutes have lapsed, say:

Now, for 5 minutes, you will need to reach a consensus with the other group members as to what you think will be the one best way of approaching the customer. You will need to reach agreement, as a group, on the one best way"

When 5 minutes have concluded, say:

"Please stop your discussion. I will now read the next scenario. Please listen carefully.

Scenario Two

A customer enters the store who looks extremely poorly dressed, their sweatshirt is ripped and they appear to be wearing no shoes. The customer is looking at items in the housewares section, and it occurs to you that they are looking around to see who might be watching them

"For 2 minutes, take some brief notes on how you think this customer should be approached. Start now."

When 2 minutes have lapsed, say:

Now, for 5 minutes, you will need to reach a consensus with the other group members as to what you think will be the one best way of approaching the customer. You will need to reach agreement, as a group, on the one best way"

"Please stop your discussion. I will now read the last scenario. Please listen carefully.

Scenario Three

You are the head salesperson in the nursery department, and at the present time, you are very busy indeed with customers and paperwork. You are so busy, you cannot possibly devote much time to customer service. A person in their teens enters the nursery department. How would you handle this situation?

"For 2 minutes, take some brief notes on how you think this customer should be approached. Start now."

When 2 minutes have lapsed, say:

Now, for 5 minutes, you will need to reach a consensus with the other group members as to what you think will be the one best way of approaching the customer. You will need to reach agreement, as a group, on the one best way"

When 5 minutes have concluded, say:

"Please stop now. Please write your name on all the material you have written on, and hand these items to the assessors".

- Step 1: Complete the behavioural checklist below, using your behavioural notes and any other material that has been completed by the participant.

Table 1. *Checklist, General Sales, Approach Simulation*

Use the following 6-point scale on the behavioural checklist for each action. Do not use fractional values:

1.	Certainly below standard
2.	Somewhat below standard
3.	Unsure, probably below standard
4.	Unsure, probably above standard
5.	Somewhat above standard
6.	Certainly above standard

Score	Expected Action
1. ☐	Suggests that the customers should be approached in some appropriate manner (Customer Focus).
2. ☐	Advocates a friendly approach to the customer, such as welcoming them, or wishing them good morning, etc. (Customer Focus).
3. ☐	Avoids business related statements on the initial approach to the customer (Customer Focus).
4. ☐	Makes suggestions that would assist to build an initial rapport with customers (Customer Focus).
5. ☐	Avoids pre-judging the customer: i.e., advocates sustaining the same level of customer service, despite the appearance of the customer (Customer Focus).
6. ☐	Writes clear and concise notes during the exercise (Comprehension).
7. ☐	Advocates that suspicious customers should be treated with a degree of caution, however, remains careful not to exert prejudice or to make accusations without evidence (Customer Focus).
8. ☐	Focuses on finding approach solutions that will fit the customer best with the other team-members (Teamwork, Tolerance).

9. ☐ Interacts in a positive and polite manner with the other participants (Teamwork, Tolerance).

10. ☐ Acknowledges and encourages other members of the group (Teamwork).

11. ☐ Speaks clearly and annunciates appropriately (Oral Expression).

12. ☐ Uses gestures and facial expressions appropriately (Oral Expression).

13. ☐ Appears to be content with the interactions in the group (i.e., the participant did not appear to become angry or frustrated in any way) (Teamwork, Tolerance).

14. ☐ Follows instructions that are given to him/her (Comprehension).

☐ From your judgement (i.e., not based on the average score) assign an *overall score* based on your perception of how well the candidate performed the actions required to complete this exercise on the behavioural 6-point scale. Do not use fractional values.

- **Step 2: Complete the competency profile rating form below using your behavioural notes and the behavioural checklist.**

Competency Profile Rating Form

2. Approach Exercise

Based on the competency profiles, and the more general scale shown below, rate the performance of the participant for each competency in the Approach Exercise. Do not use fractional values.

1. Certainly below standard
2. Somewhat below standard
3. Unsure, probably below standard
4. Unsure, probably above standard
5. Somewhat above standard
6. Certainly above standard

Competency	Rating
Comprehension	
Oral Expression	
Tolerance	
Teamwork	
Customer Focus	

Notes:

Give brief notes on the candidate's appearance:

Overall competency rating for this exercise based on your judgement (not the average), on the 1-6 scale. ***Do not use fractional values.***

Closing Exercise

(A Strategic Group Discussion Exercise)

The Closing exercise aims to elicit key characteristics that are considered important for the role of a General Salesperson. The exercise also views individual behaviour whilst interacting in a group environment.

Scenario

The Closing exercise presents a situation where the participants are given 6 different written scenarios for which they must choose appropriate ways of closing the sale as a group.

Say to the participants:

"In this exercise, you will firstly need to read through 6 different scenarios. While you are reading through them, I would like you to note down, in a clear and legible manner, how you think you should finish off the sale. In other words, ask yourself what you would say to the customer to help secure the sale of the product or products. Then you will need to take turns to individually read out a scenario, and then discuss your answers and agree on one appropriate close as a group. I will hand out these scenarios now. Please leave them face down until I tell you to read them."

Hand out scenarios face down to the participants now, and then say:

"You will soon have 8 minutes to read through the scenarios and to make some notes on how you will finish off these sales. Make sure your notes are legible because they will be assessed."

"Are there any questions at this stage? Remember, the assessors will not be able to answer any questions once the discussion has begun."

[Deal with any questions accordingly]

"You may begin reading and note taking now."

If the participants forget to take notes remind them that they will be assessed on these well in advance of 8 minute allocation.

When the participants have 2 minutes remaining, say:

"You have 2 minutes remaining."

[When 8 minutes have lapsed, say the following]

"Please stop now. You will now need to take turns to individually read out each scenario one by one. After each time a scenario is read, you will need to discuss your answers and agree on **one** appropriate close as a group. You will have 3 minutes in which to reach a consensus as a group. Note that you must reach a consensus on the one most appropriate closing strategy for each scenario"

"Are there any questions at this stage? Remember, the assessors will not be able to answer any questions once the discussion has begun."

[Deal with any questions accordingly]

"(Name of the first participant) may begin by reading out the first scenario. After the first scenario has been read, I will begin timing, and you will have a maximum of 3 minutes to reach a group consensus. Then the next participant will read the second scenario, and so on. (Name of the first participant) please read the first scenario now."

Never allow any form of voting. Only discussion leading to consensus is allowed.

After 3 minutes, say:

"Please stop your discussion. (Name of the second participant) please read the second scenario, then as a group, you must reach a consensus as to the most appropriate close for that particular situation. Please read the first scenario now."

Repeat this process until all of the scenarios have been read. At the conclusion of the exercise, say:

"Please stop what you are doing now. Please write your name on all the material you have written on, and hand these items to the assessors".

- Step 1: Complete the behavioural checklist below, using your behavioural notes and any other material that has been completed by the participant.

Table 1. *Checklist, General Sales, Closing Simulation*

Use the following 6-point scale on the behavioural checklist for each action. Do not use fractional values:

1.	Certainly below standard
2.	Somewhat below standard
3.	Unsure, probably below standard
4.	Unsure, probably above standard
5.	Somewhat above standard
6.	Certainly above standard

Score	Expected Action
1. ☐	Individually suggests closes that would be appropriate to the scenarios. (Customer Focus).
2. ☐	Assists the team by suggesting appropriate closes in a group situation (Teamwork, Customer Focus).
3. ☐	Identifies customer needs (Customer Focus).
4. ☐	Focuses on finding solutions that will assist the other team-members (Teamwork, Tolerance).
5. ☐	Interacts in a positive and polite manner with the other participants (Teamwork, Tolerance).
6. ☐	Acknowledges and encourages other members of the group (Teamwork).
7. ☐	Speaks clearly and annunciates appropriately (Oral Expression).
8. ☐	Uses gestures and facial expressions appropriately (Oral Expression).
9. ☐	Demonstrates clear annunciation when reading written passages (Oral Expression).

10. ☐ Appears to be content with the interactions in the group (i.e., the participant did not appear to become angry or frustrated in any way) (Teamwork, Tolerance).

11. ☐ Writes clear and concise notes during the exercise (Comprehension).

12. ☐ Follows instructions that are given to him/her (Comprehension).

13. ☐ Keeps a constant level of interpersonal effectiveness during the exercise (Tolerance).

☐ From your judgement (i.e., not based on the average score) assign an *overall score* based on your perception of how well the candidate performed the actions required to complete this exercise on the behavioural 6-point scale. Do not use fractional values.

- **Step 2: Complete the competency profile rating form below using your behavioural notes and the behavioural checklist.**

Competency Profile Rating Form

3. Closing Simulation

Based on the competency profiles, and the more general scale shown below, rate the performance of the participant for each competency in the Closing Simulation. Do not use fractional values.

---

1. Certainly below standard
2. Somewhat below standard
3. Unsure, probably below standard
4. Unsure, probably above standard
5. Somewhat above standard
6. Certainly above standard

Competency	Rating
Comprehension	
Oral Expression	
Tolerance	
Teamwork	
Customer Focus	

Notes:

Give brief notes on the candidate's appearance:

Overall competency rating for this exercise based on your judgement (not the average), on the 1-6 scale. Do not use fractional values.

Closing Exercise

(A Strategic Group Discussion Exercise)

The Closing exercise aims to elicit key characteristics that are considered important for the role of a General Salesperson. The exercise also views individual behaviour whilst interacting in a group environment.

Scenario

The Closing exercise presents a situation where the participants are given 6 different written scenarios for which they must choose appropriate ways of closing the sale as a group.

Say to the participants:

"In this exercise, you will firstly need to read through 6 different scenarios. While you are reading through them, I would like you to note down, in a clear and legible manner, how you think you should finish off the sale. In other words, ask yourself what you would say to the customer to help secure the sale of the product or products. Then you will need to take turns to individually read out a scenario, and then discuss your answers and agree on one appropriate close as a group. I will hand out these scenarios now. Please leave them face down until I tell you to read them."

Hand out scenarios face down to the participants now, and then say:

"You will soon have 8 minutes to read through the scenarios and to make some notes on how you will finish off these sales. Make sure your notes are legible because they will be assessed."

"Are there any questions at this stage? Remember, the assessors will not be able to answer any questions once the discussion has begun."

[Deal with any questions accordingly]

"You may begin reading and note taking now."

If the participants forget to take notes remind them that they will be assessed on these well in advance of 8 minute allocation.

When the participants have 2 minutes remaining, say:

"You have 2 minutes remaining."

[When 8 minutes have lapsed, say the following]

"Please stop now. You will now need to take turns to individually read out each scenario one by one. After each time a scenario is read, you will need to discuss your answers and agree on **one** appropriate close as a group. You will have 3 minutes in which to reach a consensus as a group. Note that you must reach a consensus on the one most appropriate closing strategy for each scenario"

"Are there any questions at this stage? Remember, the assessors will not be able to answer any questions once the discussion has begun."

[Deal with any questions accordingly]

"(Name of the first participant) may begin by reading out the first scenario. After the first scenario has been read, I will begin timing, and you will have a maximum of 3 minutes to reach a group consensus. Then the next participant will read the second scenario, and so on. (Name of the first participant) please read the first scenario now."

Never allow any form of voting. Only discussion leading to consensus is allowed.

After 3 minutes, say:

"Please stop your discussion. (Name of the second participant) please read the second scenario, then as a group, you must reach a consensus as to the most appropriate close for that particular situation. Please read the first scenario now."

Repeat this process until all of the scenarios have been read. At the conclusion of the exercise, say:

"Please stop what you are doing now. Please write your name on all the material you have written on, and hand these items to the assessors".

- Step 1: Complete the behavioural checklist below, using your behavioural notes and any other material that has been completed by the participant.

Table 1. *Checklist, General Sales, Closing Simulation*

Use the following 6-point scale on the behavioural checklist for each action. Do not use fractional values:

1.	Certainly below standard
2.	Somewhat below standard
3.	Unsure, probably below standard
4.	Unsure, probably above standard
5.	Somewhat above standard
6.	Certainly above standard

Score	Expected Action
1. ☐	Individually suggests closes that would be appropriate to the scenarios. (Customer Focus).
2. ☐	Assists the team by suggesting appropriate closes in a group situation (Teamwork, Customer Focus).
3. ☐	Identifies customer needs (Customer Focus).
4. ☐	Focuses on finding solutions that will assist the other team-members (Teamwork, Tolerance).
5. ☐	Interacts in a positive and polite manner with the other participants (Teamwork, Tolerance).
6. ☐	Acknowledges and encourages other members of the group (Teamwork).
7. ☐	Speaks clearly and annunciates appropriately (Oral Expression).
8. ☐	Uses gestures and facial expressions appropriately (Oral Expression).
9. ☐	Demonstrates clear annunciation when reading written passages (Oral Expression).

10. ☐ Appears to be content with the interactions in the group (i.e., the participant did not appear to become angry or frustrated in any way) (Teamwork, Tolerance).

11. ☐ Writes clear and concise notes during the exercise (Comprehension).

12. ☐ Follows instructions that are given to him/her (Comprehension).

13. ☐ Keeps a constant level of interpersonal effectiveness during the exercise (Tolerance).

☐ From your judgement (i.e., not based on the average score) assign an *overall score* based on your perception of how well the candidate performed the actions required to complete this exercise on the behavioural 6-point scale. Do not use fractional values.

- **Step 2: Complete the competency profile rating form below using your behavioural notes and the behavioural checklist.**

Competency Profile Rating Form

4. Closing Simulation

Based on the competency profiles, and the more general scale shown below, rate the performance of the participant for each competency in the Closing Simulation. Do not use fractional values.

1. Certainly below standard
2. Somewhat below standard
3. Unsure, probably below standard
4. Unsure, probably above standard
5. Somewhat above standard
6. Certainly above standard

Competency	Rating
Comprehension	
Oral Expression	
Tolerance	
Teamwork	
Customer Focus	

Notes:

Give brief notes on the candidate's appearance:

Overall competency rating for this exercise based on your judgement (not the average), on the 1-6 scale. Do not use fractional values.