Content validation: The forgotten step-child or a crucial step in assessment centre validation?

Authors:
Klaus-Peter Müller[1]
Gert Roodt[1]

Affiliations:
[1]Department of Industrial Psychology and People Management, University of Johannesburg, South Africa

Correspondence to:
Klaus-Peter Müller

Email:
klaus@m-network.co.za

Postal address:
PO Box 524, Auckland Park 2006, South Africa

Dates:
Received: 28 May 2013
Accepted: 19 Aug. 2013
Published: 06 Nov. 2013

How to cite this article:
Müller, K.-P., & Roodt, G. (2013). Content validation: The forgotten step-child or a crucial step in assessment centre validation? *SA Journal of Industrial Psychology/SA Tydskrif vir Bedryfsielkunde*, *39*(1), Art. #1153, 15 pages. http://dx.doi.org/10.4102/sajip.v39i1.1153

Orientation: Assessment centres (ACs) are a popular method of assessment in South Africa, as they offer a practical link to the required job, directly observed through candidate behaviour. Content is often borrowed from the USA, so research into the applicability of that content in South Africa is justified.

Research purpose: This study aimed to determine whether a selected USA-developed virtual assessment centre (VAC) measured what it claims to, and to determine whether the content is suitable for South Africa.

Motivation for the study: A solid pre-statistical foundation of content forms the backbone of assessing validity. Content validation analysis is well suited to analysing the relevance of AC simulations in a specific cultural context. Too often content validation is either implied, or insufficiently explained.

Research design, approach and method: A content evaluation schedule was developed, consisting of 50 items covering seven content validation dimensions. Thirteen subject matter experts and nine functional experts were tasked to assess an imported VAC using this schedule.

Main findings: This study provides support that the VAC appears to measure what it purports to, and that overall, the content is suitable for use in South Africa.

Practical/managerial implications: Content created in the USA can be assessed for relevance and applicability for South Africa through content validation.

Contribution/value-add: This study contributes to AC literature and assessment methodology by demonstrating the importance and utility of content validation. Importers and developers of AC content may use this study's techniques to validate content to meet legislative requirements and ensure domain relevance.

Introduction

For the past five decades, assessment centres (ACs) have globally remained a popular tool with which organisations can select, promote and train employees. Indeed, the popularity of ACs appears to be growing (Greyling, Visser & Fourie, 2003), and they are now commonly linked to strategic business processes such as talent management, leadership identification and personnel development (Schreuder & Coetzee, 2010). In South Africa, AC use has also increased since its introduction in the 1970s (Krause, Rossberger, Dowdeswell, Venter & Joubert, 2011; Meiring, 2008).

Assessment centres assess candidates on a number of job-relevant dimensions using several exercises, tasks or simulations (Jackson, Stillman & Englert, 2010). However, approximately one-third of the content used in ACs in South Africa is developed abroad and then imported for local use (Krause *et al.*, 2011). One would assume that this is due to the costs involved in the development of AC content. Gelfand (2000) mentions that over 90% of all research conducted in the field of organisational psychology is based on data from the United States and various European countries, giving the field a strong Western emphasis. This raises questions regarding the cultural fit of AC simulation content and its appropriateness for the unique diversity of South African culture. Therefore, the broad aim of this study is to establish the content validity of a virtual assessment centre (VAC) imported from the USA for use in South Africa. The virtual aspect of the VAC is due to the mainly IT-orientated method of content delivery.

Very little published research could be found on AC content validation studies. Whilst content validation has not entirely disappeared (cf. Brandt, 2005; Cohen, 1980; Grant, 2009; Rubio, Berg-Weger & Tebb, 2003; Schwartz *et al.*, 2001; Tett, Guterman, Bleier & Murphy, 2000; Van de Vijver & Tanzer, 2004), too often content validation is either implied as having been done

(Brummel, Rupp & Spain, 2009; Coetzee & Schreuder, 2009; Furtner, Rauthmann & Sachse, 2011; Golden, 1981; Lievens, 1999; McEnery & Blanchard, 1999; Roos & van Eeden, 2008; Russel & Domm, 1995; Swart, Roodt & Schepers, 1999), or no more than a paragraph is deemed sufficient explanation of content validation efforts. Preference appears to be given to construct validation studies, in which content validity is often either implied or erroneously assumed. AC construct validation studies (e.g. Furnham, Jensen & Crump, 2008; Hoffman, Melchers, Blair, Kleinmann & Ladd, 2011; Lance, Woehr & Meade, 2007; Lievens, 2009; Lievens & Thornton, 2005; Petrides, Weinstein, Chou, Furnham & Swami, 2010) are far more common than AC content validation studies. These studies only provide either a brief description of the content validation procedure or otherwise completely ignore content validation procedures. Content validation should be considered an important building block for achieving construct and criterion validity (Burton & Mazerolle, 2011).

Given the popularity of ACs in South Africa and that content is often borrowed from the USA, research on the applicability of imported AC content in South Africa is justified. This research study will provide guidance for similar assessment scenarios, and prove or disprove the efficacy of certain USA-developed VAC process content. The majority of validation research is focused primarily on construct validity; however, this is but one type of validity, and makes the assumption that the content is valid to begin with. Furthermore, it may be possible to expand the research to include possible amendments to the selected process content, case studies for example, so that content validity is increased to a satisfactory level, which may then be deemed appropriate for use in South Africa and its intended work domain. It is also possible that awareness regarding content validity and its application and limits can be created through the execution of the present study. As such, the present study will focus on one of the often neglected areas in validation research: content validity.

The results achieved through statistical procedures are significantly dependent on the quality of the data that is entered into the analysis and, as with all mathematical formulations, it is a case of: 'garbage in, garbage out'. Meaning can only be accurately inferred or deduced from data analysis if the data has relevance to begin with. This means that the relevance of content should be determined *before* statistical analysis is performed, and not *post hoc* (Haynes, Richard & Kubany, 1995). Consequently, the higher the quality of data input, the higher the likelihood of a result that is relevant and applicable to the intended domain. This argument demonstrates the importance of a high-quality content analysis.

Whilst the popularity of content validation analysis has waned within the psychological community, it has increased in the medical research community. Here, content-validated simulation exercises are a popular method of assessing and training medical staff (cf. Boulet *et al*., 2003; Curtin, Finn, Czosnowski, Whitman & Cawley, 2011; Lui *et al.*, 2010; Lauth, Magnusson, Ferrari & Petursson, 2008; McCarten & Owen, 1990; Polit, Beck & Owen, 2007; Ravens-Sieberer & Bullinger, 1998; Steadman *et al*., 2006; Tuttle *et al*., 2007). Content validation is the way in which relevance and applicability of material are assessed. Assessing and identifying the relevance and applicability of content from a domain, and determining the degree of match to a specified target population, are core functions of a content validation analysis.

The abovementioned state of affairs underscores the fact that content validation studies are either underemphasised or completely ignored. It is envisaged that this study will provide a content validation of the USA-developed virtual assessment centre under investigation.

The authors will argue and prove that a solid pre-statistical foundation of content forms the backbone in the assessment of validity. Furthermore, it will be contended that content validation analysis is well suited to analysing the relevance of AC simulations in a specific cultural context. Additionally, it will be demonstrated that content validation is an effective method with which to assess validity in its own right. If we are to prove that the content of a USA-based VAC is relevant for application in the SA context, and for the targeted job, we are required to perform a content validation study. Hence, the main research question is:

Does the selected USA-developed VAC appear to assess what it claims to, and is the content in the VAC relevant for the targeted job and appropriate for use in the South African context?

To provide supporting evidence for the research question, the following objectives have been derived:

1. To investigate to what extent experts are in agreement on the relevance of the job content domains of the USA-developed AC simulations.
2. To examine the degree of agreement between subject matter experts and experienced managers on the VAC simulation content.
3. To establish whether the USA-developed VAC simulation's content contains elements that could potentially disadvantage a particular demographic group.

Literature review

At this point, a clearer and less ambiguous definition of what an AC consists of is required. For purposes of the present study, an AC: 'consists of a standardized evaluation of behaviour based on multiple inputs' (International Task Force on Assessment Center Guidelines, 2009, p. 244). In addition, observers evaluate the behaviour under assessment by the AC simulation. Standardisation is a crucial factor in evaluation, as it dictates that all AC participants must be subjected to the same evaluation process as all other participants. A VAC uses the same fundamental base as a regular AC, with the exception that the method of content delivery is through an information technology (IT) interface. The VAC contains multiple competency-based assessment opportunities, combined in a single assessment.

The first step in assessing the relevance of a measure (a simulation exercise, for example) is performing a content

validation study. It appears that, in the South African context, particularly where ACs are concerned, content validation is performed approximately 65% of the time (Krause *et al*., 2011). However, whilst this figure may seem fairly high, it is unknown what exactly constitutes content validation efforts in ACs as little to no research has been published on the topic. The issue of assessing the validity of ACs has been hotly contested over the last 10 years, and, for the last few decades, the predominant focus has been on proving construct, predictive, and criterion validity (Hoffmann & Meade, 2012; Lievens, 1999). However, little focus has been given to the actual content that is used in ACs and perhaps content validation may be a further missing ingredient in increasing construct, predictive, and criterion validity. No matter the approach taken towards the establishment of construct, predictive and criterion validity, it will advantageous to conduct a thorough content validation beforehand.

It seems that the 'active ingredient' in ACs is the use of exercise simulations and their associated dimensions, such as conflict handling and decision-making, for example (Hoffman *et al*., 2011, p. 380). Therefore, it is logical that research on ACs should focus on the impact of simulations and their relevance to the work context. Lievens and Thornton (2005) are of the opinion that competency modelling offers numerous advantages in aligning performance in specific jobs to broader organisational objectives. They further state that organisations require certain employee competencies that enable the achievement of organisational goals. However, they point out that there is a need to develop techniques that link competencies with performance dimensions that allow competencies to be accurately assessed. The VAC is considered a modern technique that allows the assessment of a range of competencies according to performance criteria. Furnham *et al*. (2008) state that measurement dimensions are often linked to integrated competencies required by an organisation; these authors succeeded in measuring over 3000 middle level managers on seven competencies.

In the present study, the VAC was centred around and designed for middle level managerial jobs (sales managers), and was evaluated on four broad competency-related areas: decision-making, leadership, communication and management. Furthermore, dimensions of the VAC content and process were assessed according to dimensions of job competency-simulation match, as well as job complexity-simulation match, fidelity (realism), perceived demographic fairness, economic considerations and ethical considerations. Hoffman *et al*. (2011) state that ACs use different simulation exercises to measure several dimensions that are regarded as relevant to effective work performance. It is precisely this relevance that highlights the importance of performing a content validation analysis in order to determine whether a measure is relevant to a domain. The relevance of a dimension to effective work performance is established by performing a content validation on the simulation exercises and the VAC as a whole. Simulation exercises form the backbone of ACs, and are composed of some or all of the following: in-basket exercises, presentations, situational interviews, role-playing, fact finding, prioritising, decision-making, planning and leaderless group exercises.

Assessment test developers are required to follow rigorous scientific development procedures in the development of a measure. These include detailed information concerning the measure's reliability and validity. Data regarding the content validity of a measure are viewed as a necessary requirement to effectively assess possible inferences regarding the measure's quality (Polit & Beck, 2006). Furthermore, Rubio *et al*. (2003) state that, when a measure is first created, it needs to be psychometrically evaluated, and the first step is to assess the content validity of the measure.

There are numerous definitions of content validity. Content validity: 'measures the degree of correspondence between the items selected to constitute a summated scale and its conceptual definition' (Grahn & Gard, 2008, p. 70).

It: 'refers to the extent to which the items on a measure assess the same content or how well the content material was sampled in the measure' (Rubio *et al*., 2003, p. 95). Content validity is also considered to refer to 'the extent to which an instrument adequately samples the research domain of interest when attempting to measure phenomena' (Wynd, Schmidt & Schaefer, 2003, p. 509).

According to Polit and Beck (2004, p. 423) content validity is: 'the degree to which an instrument has an appropriate sample of items for the construct being measured', whilst Roodt (2009, p. 57) postulates that it: 'involves determining whether the content of a measure covers a representative sample of behaviour domain/aspect to be measured'. Content validation is also defined as: 'a multi-method, quantitative and qualitative process that is applicable to all elements of an assessment instrument' (Haynes *et al*., 1995, p. 247).

Whilst these definitions are not exactly the same, there seems to be consensus that content validity has to do with correspondence and congruency between the actual content of a measure and how representative it is of a sample of behaviours found in a specific domain. The definition of content validity used in the present study is the following:

> Content validity is the quantitative and qualitative evaluation by experts of the relevance and representativeness of an assessment measure with regard to the targeted sample domain.

In their discussion on content validity, Hoeft and Schuler (2001) mention that the founding principle of ACs is the simulation-orientated aspect of aptitude assessment. Due to the fact that AC design and construction is founded on or derived from job analysis data, it should be possible, according to Hoeft and Schuler, to: 'derive a representative image of the target position as a whole' (p. 114), which, in essence, means proving the validity of a measure through an analysis of the content.

The main principle on which a simulation exercise is based is that an AC participant is presented with situational stimuli

that require action on the part of the participant (Brummel *et al*., 2009). The simulation should be designed in such a manner that the situation provides the participant the opportunity to display desired work-related behaviour. Lievens, Tett and Schleicher (2009) mention that little is known about how simulations might be improved upon in order to more effectively elicit work-relevant behaviour. It is the author's belief that computer-based simulations are able to elicit specific actions that allow a wider range of behaviours or competencies related to a specific job to surface. The International Task Force on Assessment Center Guidelines (2009) describes a simulation as: 'an exercise or technique designed to elicit behaviors related to dimensions of performance on the job, requiring the participants to respond behaviourally to situational stimuli' (p. 246). Additionally, stimuli may be presented through various media channels, including video, audio, a computerised user interface, telephony, or any combination thereof. All of these media channels were present and supported in the VAC under investigation.

Hoffman *et al*. (2011) state that, in general, a closer inspection of ACs' construct validity is merited. However, the choice of validation method should be dependent on the AC's purpose. The purpose of the present study is not to assess construct or criterion validity, but to assess the suitability of a VAC simulation developed abroad for local implementation, through a process of content validation. One could liken content validation to evaluating the accuracy of excavating a tunnel: if one begins marginally off-centre at the beginning, the end result at the far side of the tunnel will be off by many times the original error.

Haynes *et al*. (1995) mention that the way in which content is validated should vary according to the target of the construct and the assessment method under investigation. They mention six possible steps in validating content:

1. Carefully define the domain and facets of the construct, and subject them to content validation before developing other elements of the assessment instrument.
2. Subject all the elements of an assessment instrument to content validation.
3. Use population and expert sampling for the initial generation of items and other elements.
4. Use multiple judges of content validity, and quantify judgements using formalised scaling procedures.
5. Examine the proportional representation of items.
6. Report results on content validation when publishing a new assessment instrument.
7. Use subsequent psychometric analyses for refinement of the assessment instrument.

If content is created abroad, how are we to know if it is applicable and amendable for implementation in South Africa? Firstly, a clear definition of the domain or competency area and all its facets is required. Secondly, the appropriateness and relatedness of the content has to be assessed: expert judges have to assess the relevance of content and its cross-cultural applicability. Furthermore, expert judgement should be used in the assessment of content quality. According to Sackett (1987), expert judgement should consider the following dimensions and facets in the evaluation of content validity: dimensions must be shown to be related to important job activities, exercises must be shown to represent the most common and most significant job activities and must also reflect the complexity and difficulty of job activities, and dimensions must be shown to be observable in the exercises. Evidence suggests that the higher the degree of content validity underpinning an assessment method, the higher the probability that the assessment will have a high level of criterion-related validity (Sproule, 2009).

Cultural equivalence of measurements is an important area in assessment research. Content that is valid in one culture is not automatically valid or appropriate in another. Thus, taking assessment measures and applying them in a different cultural context other than the one in which they were created, without testing for equivalence or fairness, invariably leads to concerns regarding unfairness, bias and, most importantly, inaccurate decision-making. The term *adverse impact* is used to describe the scenario in which selection procedures have varying selection rates for different cultures, genders and ages (Risavy & Hausdorf, 2011). This may be a further possible consequence of an assessment measure's content bias. The negative consequences of Western quantitative research conducted in non-Western cultures have been well documented, especially with regard to studies on cognitive abilities (Bodkin-Andres, O'Rourke, Grant, Denson & Craven, 2010; De Corte, Sackett & Lievens, 2010). However, the practical, strong job task-related nature of AC simulations should assist in reducing this, as cognitive ability is usually one of several areas under assessment. Additionally, Thornton and Gibbons (2009) mention that historically ACs show little evidence of systematic bias.

Limited literature is available on the suitability of USA-developed AC content for use in South Africa. This may be due to the direct assessment nature of ACs, which requires participants to demonstrate work-specific competencies, and the presumed universality of job tasks around the world. There exists, however, empirical research on the cross-cultural adaptation and validation of psychometric instruments for use in South Africa (cf. Byrne & Van de Vijver, 2010; De Bruin & Buchner, 2010; De Klerk, Boshoff & Van Wyk, 2010; Edwards & Leger, 1995; Gradige & De Jager, 2011; Marais, Mostert & Rothmann, 2009; Mosdell, Balchin & Ameen, 2010; Moyo & Theron, 2011; Olckers, Buys & Grobler, 2010; Oosthuizen & Koortzen, 2009; Rothmann, Mostert & Strydom, 2006; Storm & Rothmann, 2003; Twigge, Theron, Steel & Meiring, 2005; Visser & Viviers, 2010). Issues of bias and fairness in testing have received strong focus, not only from the political arena, but also from the academia. It is crucial to minimise negative impact on demographic groups in order to advance fairness in personnel decisions.

Unfairness and bias can be defined as the incomparability of samples regarding aspects other than the target variable (Van

de Vijver & Tanzer, 2004). The goal of research in this area is to ensure either complete or a high degree of equivalence of measures within a specific cultural context. Furthermore, Van De Vijver and Tanzer (2004) describe cultural equivalence as the absence of bias, which is a prerequisite for valid comparisons across cultural populations. As far as is possible under the current design, the VAC content under investigation will be assessed for aspects in the content that could potentially disadvantage a particular demographic group.

Fairness and accuracy are of paramount importance in decision-making regarding the selection, promotion and training of personnel. This fairness is not only ethical, it is legislated in the *Employment Equity Act* (Republic of South Africa, 1998). Making decisions regarding employees without accurate, scientifically valid information would be random, at best. Additionally, personnel decisions have taken on a more strategic role in organisations, evident in practices such as talent management and leadership development. It therefore stands to reason that incorrect decisions will have grave negative consequences, and possibly even threaten the organisation's survival. South Africa is hyper-vigilant in ensuring that employees are not adversely impacted by the use of non-scientific, invalid and unreliable psychological measures.

The present study focuses on assessing an existing AC and, as such, was not part of the initial creation and design process. Additionally, the work domain that is the focus of the AC had been previously defined by a generic job competency modelling framework. The VAC is not a classic psychometric assessment instrument; it is a collection of simulations where observable behaviour is assessed according to job-specific criteria. Assessment of the content validity of the AC in the present study will involve several subject matter experts, who evaluate and judge the AC's content on:

1. The relevance of a particular competency to the specified job.
2. The representativeness of the VAC simulations with regard to the specified job.
3. The comprehensiveness with which the VAC assesses competencies.
4. Elements that could disadvantage a particular demographic group.

The perceptions and opinions of the experts on a defined domain or competency area will aid in proving the content validity of the VAC.

Research design

Research approach

Various stakeholders with a variety of perceptions regarding the AC process were involved in assessing the content of an AC. These perceptions are important in determining the relevance and appropriateness of the content of the AC. The main stakeholders in the assessment of content validity in the AC process are: human resource (HR) specialists, participants, subordinates and managers (Roodt, 2008). The present study focused on the perceptions of two of these stakeholder groups, namely managers and HR specialists. It is the present author's view that managers and HR specialists could be expected to accurately decide on the relevance and applicability of the VAC content. According to Roodt (2008), HR specialists contribute to the content evaluation of an AC with regard to overall competency profile match. Roodt furthermore states that managers can provide beneficial data on the content relevance and accuracy of an AC process.

Content analysis and, especially, the manner in which the authors wish to use it to infer content validity, require the move from a purely descriptive analysis towards an empirical analysis, which allows meaningful inferences to be drawn from the content. Rourke and Anderson (2004) propose the following steps in developing a theoretically valid protocol for establishing content validity:

1. Identify the purpose of the coding data.
2. Identify behaviours that represent the construct.
3. Review the categories and indicators.
4. Hold preliminary try-outs.
5. Develop guidelines for administration, scoring and interpretation of the coding scheme.

It is assumed that this process was meticulously followed by the AC developers in developing the VAC content under investigation.

Content analysis, in the present context, establishes how well or how poorly the VAC process content corresponds to actual job content. It is critical to identify behavioural competencies that represent the construct to be measured by the VAC simulations, as the simulation exercises are designed to tap into the identified behaviours or competencies. Simulation exercises are typically derived from a job analysis or a competency-based job analysis. The latter was used in the present study. A structured evaluation schedule was developed, which provided a competency-informed context, covering information that was representative of the target sample (sales managers). The subject matter experts were then tasked with rating their agreement or disagreement with the content being related to the target sample in terms of relevance, clarity, simplicity and ambiguity (Yagmaie, 2003).

Research method

Participants

A non-random purposive sampling approach was used in the selection of participants, since it is desirable to source participants whose experience and knowledge relate to a specific topic, which, in the present study, included ACs and managerial-related competencies. Morse (1994) postulates the following criteria for obtaining a participant who is suitable to a study: the participant must have knowledge and experience of the phenomenon under scrutiny, they must have the capacity to express themselves, and must be willing to participate in the study. Purposive sampling is believed to enhance the trustworthiness of a study (Groenewald, 2004).

Additional criteria were set for the inclusion of a participant in the present study, which related to their qualifications: a participant had to be either a registered industrial psychologist with at least two years' experience in ACs, hereafter referred to as subject matter experts (SMEs), or they had to be a manager with intimate knowledge of the job under assessment. These managers are referred to as functional experts (FEs). A total of 22 (N = 22) participants were purposively selected from the population of South Africa's Gauteng region, who then voluntarily chose to take part in the research study. Demographic data such as age, gender, language and expert status were elicited from the participants. Table 1 reflects the characteristics of the participants.

Table 1 indicates that approximately 23% of the participants were aged 25–34 years, whilst nearly 41% were aged 35–44. About 18% were aged 45–54 years, and 18% were aged 55–64 years. Seventy-six percent of the participants were above the age of 34. An even number of men and women took part in the study (50% respectively). The participants were predominantly English speaking (nearly 55%), whilst about 41% were Afrikaans speaking, and 4% were IsiXhosa speaking. Approximately 59% of the participants were classified as SMEs, whilst 41% were classified as FEs.

TABLE 1: Characteristics of participants.

Item	Category	SMEs		FEs		Combined	
		f	%	*f*	%	*f*	%
Age	25–34	1	7	4	44.4	5	22.7
	35–44	4	30.7	5	55.6	9	40.9
	45–54	4	30.7	0	0	4	18.2
	55–64	4	30.7	0	0	4	18.2
Gender	Male	5	38	6	66.6	11	50
	Female	8	52	3	33.3	11	50
Language	English	4	30.7	8	88.8	12	54.5
	Afrikaans	8	52	1	11.2	9	40.9
	IsiXhosa	1	0.7	0	0	1	4.5
Total	-	**13**	-	**9**	-	**22**	**100**

SME, subject matter experts; FE, functional experts; *f*, frequency.

Measuring instrument

Data were collected via a structured evaluation schedule. The evaluation schedule contained both quantitative and qualitative items (although the majority were quantitatively orientated). To ascertain the level of agreement on the quantitative section, a Likert-type intensity scale ranging from 1–7 was used. A space for comments was provided after each dimension rating section, where experts could add information that they considered relevant and important to the dimension being evaluated. The evaluation schedule was informed by aspects identified in the literature review, the job competency profile and the characteristics of the intended work domain.

Elements of the work domain must be assessed for relevance, as recommended by Rubio *et al*. (2003), who suggest rating a domain on a scale of 1–4 for relevance and clarity. However, the present study made use of a seven-point Likert interval-based scale, as it was determined that distances within item scores had meaning, and the experts were considered able to distinguish between item increments. A lower bound score of 1 indicated the complete absence of a property whereas a higher bound score of 7 indicated the definite presence of a property. A score of 4 indicated the scale mid-point. Rubio et al. also recommend evaluating the clarity of the AC together with each item used to measure the representativeness of the VAC. This allows experts to evaluate both dimensions simultaneously, negating the need to remember information previously provided on another dimension. Furthermore, multiple dimensions are generally listed for a construct under inspection; if dimensions are present in the measure, experts can be tasked with assigning the item to its relevant construct. The experts were also asked to report on areas that they deemed important, but which were not found in the measurement instrument, and to do so in the space provided for comments.

Seven dimensions were identified from the literature and assessed using the evaluation schedule. The dimensions identified are as follow below.

Competency area-job correspondence: This dimension comprised five facets relating to the degree of competency sampling as targeted by the VAC. Competencies included: critical thinking, leadership, communication, task process management and client focus. These facets were included in the evaluation schedule in order to establish the degree to which the job content and the job competencies corresponded. Managerial competencies such as critical thinking and leadership were assessed with items such as: 'To what extent does the VAC assess analytical (critical thinking and decision-making) behaviour?' and 'To what extent is people leadership effectively sampled by the VAC?' The sub-dimensions of communication, task process management and client orientation were assessed with items such as 'How comprehensive is the VAC at sampling communication?', 'How diverse is the range of managerial-type (task and process) behaviour elicited by the VAC?' and 'To what extent are client-focused skills effectively sampled by the VAC?'

Job competency-simulation match: This dimension was assessed to determine the extent to which the VAC simulations match the job competencies. Congruency between the VAC simulations and related job competencies was the desired outcome for this portion of the assessment. Items such as 'All competency areas considered, how comprehensively does the VAC match the job profile?' were used to assess overall match.

Complexity: When addressing the dimension of complexity, it is important to demonstrate that the complexity level of the VAC matches the job complexity. This match was assessed with items such as 'How appropriate is the level of difficulty to a VAC participant who is at a managerial level?'

Fidelity: It was important to assess and compare the realism of the VAC simulations to a real-life, work-related scenario.

Items such as 'How typical are the problems encountered in the VAC compared to problems found in the actual job?' were used to assess the fidelity construct.

Potential demographic fairness: Here, the authors were attempting to establish potential issues in the content that could lead to a demographic group being disadvantaged. It was thought that if a group was disadvantaged, this could lead to poor performance in the VAC which would then in turn influence evaluation scores. Therefore, it was vital to include a rating dimension that took into consideration issues in the content that could possibly disadvantage a particular demographic group. Items such as 'To what extent did the VAC raise concerns regarding gender discrimination?' and 'How biased is the VAC in terms of a participant's race?' were used to assess fairness in the content of the VAC.

Economic considerations: Here, the potential value and effectiveness of the VAC were assessed in order to determine the perceived relative economic value of the VAC. This was done to determine if the VAC is redundant when compared with other means of assessing a participant's job eligibility. Items such as 'How effective is the VAC in eliciting competencies in an efficient (timely and accurate) manner?' and 'How likely is it that other methods of assessment would generate similar levels of information in the allocated time?' were used to assess economic value considerations.

Ethical considerations

Here, possible and potential ethical issues of the VAC were addressed. Simply put, ethics is concerned with balancing the interests of the self and the other (Van Vuuren, 2010). It was important to determine whether the VAC encourages good conduct in VAC candidates. Items such as 'To what extent does the VAC require participants to act against socially accepted norms?' were used to assess implicit and explicit potential ethical issues in the VAC.

Research procedure

The research procedure that was implemented was informed by best practices (cf. Haynes *et al*., 1995; Kolk, Born, Van der Flier & Olman, 2002; Sackett, 1987; Sproule, 2009; Yagmaie, 2003), and guided by procedures from content validation (cf. Polit, Beck & Owen, 2007; Rubio *et al*., 2003) and cross-cultural adaptation studies (cf. Brandt, 2005; Mosdell, Balchin & Ameen, 2010; Van de Vijver & Tanzer, 2004). Additionally, issues regarding perceived demographic fairness were informed and guided by practices and procedures adapted from previous studies (cf. Bosco & Allen, in press; De Corte *et al*., 2010; Risavy & Hausdorf, 2011). As there is no standardised procedure for content validation, the abovementioned practices and procedures were modified to suit the present content validation study.

Experts were tasked with rating VAC process content that the candidate encountered according to the abovementioned criteria set out in the evaluation schedule. The actual candidate was not rated; only the VAC process content was rated.

Figure 1 reflects how the two types of experts evaluated the same content process. The role of expert raters, as illustrated in Figure 1, was *not* to evaluate or score an individual candidate's performance on the VAC, but to observe and evaluate the candidate interfacing with the content during the VAC.

Figure 1 shows how the two types of experts viewed the same virtual assessment process. The experts viewed and experienced the assessment via a computer interface with a Web browser, as well as telephonically. Both the SMEs and the FEs observed a VAC candidate completing the VAC with a view to assessing the content that the candidate encountered. The VAC contains various mini-simulations, such as the case of an irate customer, during which conflict management skills are observed. After the VAC was completed, expert raters (both SMEs and FEs) were tasked with completing the VAC evaluation schedule. In total, 22 experts scored 50 items relating to the seven abovementioned dimensions.

Statistical analysis

Kottner *et al*. (2011) mention that there are several statistical approaches that may be implemented in the assessment of agreement, but that no single approach is regarded as standard. Perhaps, this is partly due to the nature of expert rater design studies, in which the level of agreement between a set of particular raters on a particular instrument is assessed at a particular time. This results in agreement being a property of the testing context and not of the instrument itself (Hallgren, 2012). Stemler (2004) mentions that there are three broad methods through which inter-rater reliability may be established: consensus estimates (Cohen's Kappa, Kendall's *W*), consistency estimates (Intraclass correlation coefficient, Pearson correlation coefficient), and measurement estimates (factor or Rasch analysis).

Over the last few years, various attempts have been made to find a single index of content validity using rater agreement. These have yielded, for example, the Content Validity Index (CVI) (Polit, Beck & Owen, 2007). In another attempt at finding a single index, the Factorial Validity Index (FVI), as used by Rubio *et al*. (2003), was considered as a possible method with which to assess the positive matching of dimensions or factors by experts in the present study.

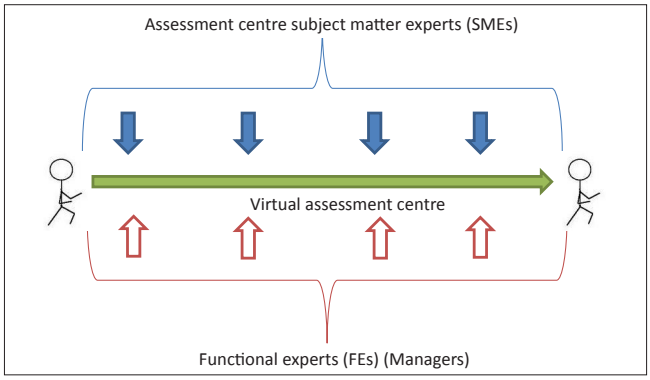

FIGURE 1: The role of subject matter experts and functional experts in evaluating the virtual assessment centres process.

However, Stemler (2004) mentions that attempts at using a single unified concept to assess agreement amongst raters is at best imprecise and at worst potentially misleading. It was decided not to use the CVI or the FVI, due to their sensitivity when more than five raters are involved.

Perreault and Leigh (1989) focused on the reliability of nominal data based on qualitative judgments, and state that the quality of judgement-based nominal scale data is often neglected and underestimated in research. They developed a parsimonious approach to estimating reliability (Ir), using expert judges and based on Cohen's kappa coefficient. However, as the present study made use of multiple raters evaluating the same process, the intraclass correlation coefficient (ICC) was deemed more appropriate. To assess inter-rater reliability (agreement), Shrout and Fleiss (1979) recommend using ICC, in which n targets are rated by k judges. The ICC is used to assess inter-rater reliability amongst raters by comparing the variability of different ratings of an identical subject or process to the total variation across all ratings and subjects (Uebersax, 2007). The terms inter-rater agreement and inter-rater reliability are used interchangeably in this study. Hallgren (2012) states that in cases where it is important that raters' scores are similar in rank order that a Cronbach's alpha consistency model should be used. An advantage with a consistency estimate such as Cronbach's alpha is that it allows for an overall estimate of consistency amongst multiple judges (Stemler, 2004).

Inter-rater expert agreement is the main method used to assess the validity of content, and has been used in many empirical studies to successfully analyse content validity (Bakker, Van Emmerik & Van Riet, 2008; Brummel *et al*., 2009; Chung, Chiang, Chou, Chu & Chang, 2010; Kolk, Born, Van der Flier & Olman, 2002; Kunz, 2010; Mokkink *et al*., 2010; Polit et al., 2007; Rein *et al*., 2011; Wright & Craig, 2011). It should be noted that the majority of the aforementioned studies analysed content at a measurement scale item level whereas the current study analysed the content of an entire process.

Measurement estimates such as factor analysis and the many facets Rasch model (Bond & Fox, 2007) are based upon the notion that all available data from raters should be used when deriving a cumulative score for each rater (Multon, 2010; Stemler, 2004). Uebersax (1988) proposed a probability model related to latent class analysis to allow inferences regarding validity to be made from rater scores or evaluations. Furthermore, Uebersax and Grove (1993, p. 833) state that: 'the latent trait approach evaluates rating precision based on agreement with a *latent consensus*'. Stemler (2004) states that, through the application of factor analytic methods, multiple judges may rate a set of participants. Furthermore, stating that this method will allow the determination of the amount of shared variance in the ratings. Uebersax and Grove recommend using a maximum likelihood estimation to produce parameter estimates, which can be used to evaluate multiple rater performance.

It is imperative that the reader understand that it is expert agreement of AC content that is being assessed and not psychometric constructs. When use is made of parametric statistical methods such as factor analysis, it is not for the identification or categorisation of factors as is the case when identifying constructs such as conscientiousness from the Basic Traits Inventory (De Bruin & Rudnick, 2007). Rather, it is to assess the relationship (correlation) between expert's opinion scores and a common shared factor (VAC process, for example). The premise is that all raters should be scoring the same common event, process, or person. A further issue that needs to be addressed is that of statistical power. Three things are generally required to calculate statistical power: statistical significance, sample size and effect size (Hair, Black, Babbin & Anderson, 2010). In the present study, no hypotheses were tested and no consecutive administrations were carried out which means that there were no effect sizes present. Thus, it is impossible with the current research design to assess for statistical power. Results of this study are not intended for generalisability. However, it is possible to calculate statistical power and achieve a level of generalisability using an inter-rater test-retest reliability design (Shoukri, Asyali & Donner, 2004).

To determine a degree of content validity from expert ratings of the VAC content, three supporting pieces of evidence are needed. Firstly, that experts scored identified dimensions highly. Secondly, that expert scores show a high degree of consistency (inter-rater agreement) with one another. Thirdly, an estimate that shows the level to which experts were viewing and rating the same process as each other. In the present study, statistical procedures were selected that would generate a representative perspective of data that, when combined, will provide supporting evidence of the existence or nonexistence of content validity.

Data analysis was performed using IBM's SPSS (version 20.0) statistical program package. Descriptive statistics – means and standard deviations – were calculated to describe the data. The reliability of the evaluation schedule was assessed by means of Cronbach's alpha coefficients. Next, the ICC coefficient (Cronbach's alpha consistency model) was computed to determine levels of inter-rater agreement. Additionally, the ICC was used to assess rating reliability by comparing the variability of different ratings of the same subject to the total variation across all ratings and all subjects. In plain terms, it was used to assess the consistency of rater scores with one another. Factor loadings for raters on a single factor were calculated to provide an indication that raters are rating the same construct and to assess the extent that raters have consensus (Stemler, 2004). This is the equivalent to the correlations of each raters' ratings with the single common extracted factor. The Mann-Whitney U test for independent samples was used to determine whether group differences existed. Z-scores were calculated to establish whether the group scores of the SMEs and the FEs on dimensions were both within one standard deviation score to determine how close group scores clustered together around the mean and to each other.

Results

Table 2a and Table 2b show descriptive statistics for the seven dimensions. The first is Competency area– job correspondence (CAJC), which is comprised of the following five facets: critical thinking (CT), people leadership (PL), communication (COMM), task process management (TPM) and client focus (CFOC). The remaining dimensions are job competency – simulation match (JCM), job complexity – simulation match (CMPLX), fidelity (FDL), potential demographic fairness (PDF), economic considerations (ECO) and ethical considerations (ETH).

In Table 2a and Table 2b, the mean scores and descriptive statistics for the dimensions are shown. Means (with standard deviations in parentheses) for dimensions Critical thinking through ETH were 5.69 (0.85), 4.98 (1.08), 5.42 (1.12), 5.41 (0.91), 5.44 (1.24), 5.53 (0.84), 5.4 (0.77), 5.59 (1.07), 6.49 (0.59), 4.8 (0.75), 6.34 (0.79) respectively. These mean scores are somewhat higher than the scale's mid-point (4), which suggests slightly negatively skewed score distribution curves. Average scores for PDF (6.49) and ETH (6.34) were particularly high. Overall, the majority of mean scores for dimensions are high, with relatively small standard deviations. This can be taken to indicate that most agreement amongst raters on the dimensions relating to VAC content validity is high. Overall, a Cronbach's alpha coefficient of .96 was recorded for the evaluation schedule. This indicates the high degree (α > .80) of overall reliability (internal consistency) of the evaluation schedule.

On a dimension level, the following Cronbach's alpha coefficients were found: CT (.82), PL (.92), COMM (.9), TPM (.91), CFOC (.96), JCM (.86), CMPLX (.74), FDL (.92), PDF (.82), ECO (.65) and ETH (.41). The last two dimensions had low reliabilities.

Table 3a and Table 3b reflect the ICCs for experts (SMEs and FEs) on the evaluation schedule dimensions.

Table 3a and Table 3b show that the ICC average alpha consistency coefficients for both SMEs and FEs are largely congruent with each other. However, visible differences in ICC alpha scores between the SME and FE groups respectively for CMPLX (α = .79; α = .69) and PDF (α = .64; α = .98) were found. The low negative ICC value for SME on the ETH (α = –.15) dimension is due to the negative average covariance amongst items. According to Cicchetti (1994),

TABLE 2a: Summary of means, confidence intervals, variances, standard deviations, minimum and maximum scores of schedule dimensions.

Dimension	$\bar{x}$	95% CI		S^2	SD	Minimum	Maximum
		LL	UL				
CAJC							
CT	5.69	5.32	6.07	0.72	0.85	3.67	7
PL	4.98	4.5	5.46	1.17	1.08	2.33	7
COMM	5.42	4.93	5.92	1.25	1.12	3	7
TPM	5.41	5	5.81	0.84	0.91	3.67	7
CFOC	5.44	4.89	5.99	1.53	1.24	3	7

$\bar{x}$, mean; CI, confidence interval; LL, lower limit; UL, upper limit; S^2, variance; SD, standard deviation; CAJC, Competency area – job correspondence; CT, critical thinking; PL, people leadership; COMM, communication; TPM, task process management; CFOC, client focus.

TABLE 2b: Summary of means, confidence intervals, variances, standard deviations, minimum and maximum scores of schedule dimensions.

Dimension	$\bar{x}$	95% CI		S^2	SD	Minimum	Maximum
		LL	UL				
JCM	5.53	5.16	5.9	0.7	0.84	4	6.8
CMPLX	5.4	5.06	5.74	0.59	0.77	3.6	6.6
FDL	5.59	5.59	6.07	1.16	1.07	3.14	7
PDF	6.49	6.23	6.71	0.35	0.59	5.14	7
ECO	4.8	4.47	5.14	0.57	0.75	3.29	6.43
ETH	6.34	5.99	6.69	0.63	0.79	4.75	7

$\bar{x}$, mean; CI, confidence interval; LL, lower limit; UL, upper limit; S^2, variance; SD, standard deviation; JCM, job competency – simulation match; CMPLX, job complexity – simulation match; FDL, fidelity; PDF, potential demographic fairness; ECO, economic considerations; ETH, ethical considerations.

TABLE 3a: Inter-rater reliability (ICC) of expert ratings on schedule dimensions.

Dimension	Expert	ICC		*f*	*df*	*p*
		Single α	Average α			
CAJC						
CT	SME	.54	.78	4.49	12	.01
	FE	.59	.81	5.33	8	.01
	Overall	.61	.82	5.63	21	.01
PL	SME	.68	.87	7.503	12	.01
	FE	.86	.95	19.19	8	.01
	Overall	.78	.92	11.88	21	.01
COMM	SME	.66	.86	6.92	12	.01
	FE	.85	.94	17.70	8	.01
	Overall	.79	.90	9.88	21	.01
TPM	SME	.66	.87	6.93	12	.01
	FE	.80	.92	13.00	8	.01
	Overall	.77	.91	11.16	21	.01
CFOC	SME	.80	.92	13.14	12	.01
	FE	.89	.96	25.71	8	.01
Overall		**.89**	**.96**	**25.31**	**21**	**.01**

CAJC, Competency area – job correspondence; CT, critical thinking; PL, people leadership; COMM, communication; TPM, task process management; CFOC, client focus; SME, subject matter experts; FE, functional experts; *f*, frequency; *df*, degrees of freedom; *p*, probability.

TABLE 3b: Inter-rater reliability (ICC) of expert ratings on schedule dimensions.

Dimension	Expert	ICC		*f*	*df*	*f*
		Single α	Average α			
JCM	SME	.50	.83	6.02	12	.01
	FE	.45	.81	5.16	8	.01
	Overall	.55	.86	7.02	21	.01
CMPLX	SME	.43	.79	4.70	12	.01
	FE	.30	.69	3.17	8	.01
	Overall	.36	.74	3.77	21	.01
FDL	SME	.46	.86	6.91	12	.01
	FE	.60	.91	11.32	8	.01
	Overall	.63	.92	12.95	21	.01
PDF	SME	.20	.64	5.41	12	.01
	FE	.60	.98	11.60	8	.01
	Overall	.38	.82	5.41	21	.01
ECO	SME	.21	.64	2.81	12	.01
	FE	.16	.58	2.35	8	.03
	Overall	.21	.65	2.88	21	.01
ETH	SME	-.03	-.15	0.87	12	.58
	FE	.49	.77	4.38	8	.01
	Overall	.15	.41	1.69	21	.57
Overall		**.35**	**.96**	**24.95**	**21**	**.01**

JCM, job competency – simulation match; CMPLX, job complexity – simulation match; FDL, fidelity; PDF, potential demographic fairness; ECO, economic considerations; ETH, ethical considerations; SME, subject matter experts; FE, functional experts; *f*, frequency; *df*, degrees of freedom; *p*, probability.
ETH SME value negative due to negative average covariance amongst items.

ICC values of agreement falling within the range .60 and .74 can be considered as good and values of between .75 and 1.0 considered excellent. Apart from ETH (average α = .41) and ECO (average α = .65), the average ICC alpha consistency scores ranged from .74 to .96.

Single ICC is an index for the reliability of the ratings for one, typical (average), single rater. FEs typically scored, on average, higher on dimension reliability than their SME counterparts. Large differences were found in single ICCs between SMEs' and Fes' values for COMM (α = .66, α = .85), TPM (α = .66; α = .80) and PDF (α = .20; α = .60).

The one single shared event that expert raters had in common is that which they were rating. According to the design of this study, all expert raters should be observing and scoring the same process content. As such, a single factor is extracted from data to reflect the commonness of agreement in relation to the shared event. Technically, this is similar to the manner in which item loadings are analysed, the factor loadings of each rater (as opposed to item) on a factor is analysed.

Table 4 displays the communalities and factor loadings of each rater on a single factor, and the correlation coefficient of each rater with the group score.

In Table 4, it is shown that a single factor was extracted using a maximum likelihood estimation as recommended by Uebersax and Grove (1993), and that the result is interpretable (χ^2 (209) = 400.97; $p \leq .01$). Overall, approximately half the raters' communalities were above .40. Extremely low communalities were found for Rater 6 (.03), Rater 11 and Rater 15 (.05). Similarly, for the same raters, low factor loadings of between .18 and 22 were found. Stemler (2004) states that factor loadings greater than .60 indicate that raters are scoring a common construct. Multon (2010) suggests factor loadings of greater than .70, but this seemsexcessively high and appears to be more of an ideal value to strive towards.

The correlation between a single rater and the entire group shows how much each single rater corresponded with the group. Single rater correlation with the overall rater group score values ranged from .19 to .49 on the different rated dimensions. Comparatively, fairly low to moderate correlations were found for Rater 3 (.19), Rater 11 (.22), Rater 15 (.21) and Rater 20 (.25). Sixty-eight percent of rater group correlations were greater than or equal to .30.

Table 5a and Table 5b show the Mann-Whitney U test between two independent samples for SMEs and FEs.

From Table 5a and Table 5b, it can be concluded that there are statistically significant differences between the SMEs (n = 13) and FEs (n = 9) in terms of TPM (U = 24; p = .02), CFOC (U = 24.5; p = .02), JCM (U = 23.5; p = .02), FDL (U = 22.5; p = .02) and ECO (U = 26; p = .03). A further analysis of SMEs' and FEs' z-scores is shown in Table 6a and Table 6b.

From Table 6a and Table 6b, a constant universal difference between the z-scores for SMEs and FEs on all dimensions can be observed. Absolute values on dimensions CT (0.85), PL (0.83), COM (0.47), CMPLX (0.54), PDF (0.25), ECO (0.86) and ETH (0.63) were all within the range of 1 Standard Deviation from each other, indicating a good degree of similarity amongst SME and FE scores. However, for dimensions TPM (1.06), CFOC (1.23), JCM (1.18) and FDL (1.33), the z-scores differences fell outside the range of 1 Standard Deviation. This can be interpreted as moderate to slight differences on the preceding dimensions with regard to SMEs and FEs.

To infer a level of VAC content validity, it is firstly required that raters rate relevant dimensions highly. A mean score of

TABLE 4: Rater communalities, factor loadings, and single rater vs. group correlation.

Rater	Communalities	Factor Loadings	Rater vs. Group Correlation
R1	.65	.81	.45
R2	.51	.71	.41
R3	.22	.47	.27
R4	.74	.86	.46
R5	.30	.54	.32
R6	.03	.18	.19
R7	.58	.76	.41
R8	.77	.88	.49
R9	.18	.42	.30
R10	.44	.66	.42
R11	.05	.21	.22
R12	.23	.47	.33
R13	.50	.71	.43
R14	.22	.47	.29
R15	.05	.22	.21
R16	.75	.86	.48
R17	.37	.61	.32
R18	.78	.88	.47
R19	.28	.52	.35
R20	.18	.42	.25
R21	.30	.54	.35
R22	.13	.36	.26

Extraction method: Maximum likelihood. χ^2 (209) = 400.97, Significant $p \leq .01$

TABLE 5a: Mann-Whitney U test of subject matter experts (n = 13) and functional experts (n = 9) differences on evaluation schedule dimensions.

Dimension	Mann-Whitney U	Wilcoxon W	z	p
CAJC				
CT	34.00	79.00	-1.67	.09
PL	30.50	75.50	-1.89	.06
COMM	43.50	88.50	-1.01	.31
TPM	24.00	69.00	-2.33	.02*
CFOC	24.50	69.50	-2.29	.02*

CAJC, Competency area – job correspondence; CT, critical thinking; PL, people leadership; COMM, communication; TPM, task process management; CFOC, client focus; z, z-score.
*, $p \leq .05$ (2-tailed)

TABLE 5b: Mann-Whitney U test of subject matter experts (n = 13) and functional experts (n = 9) differences on evaluation schedule dimensions.

Dimension	Mann-Whitney U	Wilcoxon W	z	p
JCM	23.50	68.50	-2.39	.02*
CMPLX	44.00	89.00	-0.98	.33
FDL	22.50	67.50	-2.41	.02*
PDF	57.50	148.50	-0.07	.95
ECO	26.00	71.00	-2.18	.03*
ETH	42.50	133.50	-1.11	.27

JCM, job competency – simulation match; CMPLX, job complexity – simulation match; FDL, fidelity; PDF, potential demographic fairness; ECO, economic considerations; ETH, ethical considerations; z, z-score.
*, $p \leq .05$ (2-tailed)

TABLE 6a: Z-Scores of subject matter (SME) and functional experts (FE) on dimensions on evaluation schedule.

Dimension	$\bar{x}$ SME	$\bar{x}$ FE	($\bar{x}$ SME + $\bar{x}$ FE) / 2	SD (SME)	SD (FE)	Z-SME	Z-FE	|Z-SME| +|Z-FE|
CAJC								
CT	5.97	5.3	5.63	0.66	0.96	0.51	-0.34	0.85
PL	5.33	4.48	4.90	0.89	1.17	0.47	-0.36	0.83
COM	5.64	5.11	5.37	0.98	1.28	0.27	-0.20	0.47
TPM	5.77	4.89	5.33	0.79	0.85	0.55	-0.51	1.06
CFOC	5.97	4.67	5.32	0.89	1.29	0.73	-0.50	1.23

$\bar{x}$, mean; SD, standard deviation; *Z*, *Z*-score; CAJC, Competency area – job correspondence; CT, critical thinking; PL, people leadership; COMM, communication; TPM, task process management; CFOC, client focus.

TABLE 6b: Z-Scores of subject matter (SME) and functional experts (FE) on dimensions on evaluation schedule.

Dimension	$\bar{x}$ *SME*	$\bar{x}$ *FE*	($\bar{x}$ *SME* + $\bar{x}$ *FE) / 2*	SD (SME)	SD (FE)	Z-SME	Z-FE	*|Z-SME| +|Z-FE|*
JCM	5.88	5.02	5.45	0.63	0.86	0.68	-0.5	1.18
CMPLX	5.57	5.16	5.36	0.68	0.85	0.30	-0.24	0.54
FDL	6.08	4.89	5.48	0.72	1.15	0.82	-0.51	1.33
PDF	6.55	6.4	6.47	0.46	0.76	0.16	-0.09	0.25
ECO	5.05	4.44	4.74	0.67	0.74	0.45	-0.41	0.86
ETH	6.54	6.06	6.3	0.62	0.95	0.38	-0.25	0.63

$\bar{x}$, mean; SD, standard deviation; Z, Z-score JCM, job competency – simulation match; CMPLX, job complexity – simulation match; FDL, fidelity; PDF, potential demographic fairness; ECO, economic considerations; ETH, ethical considerations.

5 (out of 7) and larger is considered a high score. Overall, the majority of mean scores for schedule dimensions met this criterion, with relatively small standard deviations.

Secondly, it is required that raters' scores show a degree of consistency with one another. A high level of inter-rater reliability was found across all dimensions (CT, PL, COMM, TPM, CFOC, JCM, CMPLX, FDL and PDF), with the exception of dimensions ECO (economic considerations) and ETH (ethical considerations). Regarding ECO, the low inter-rater reliability could be taken to mean that the items constituting this dimension could be further refined to aid clearer understanding. Additionally, the high level of dimension reliability and consistency indicates that raters' scores show a high level of rater dimension agreement There seems to be little consensus amongst participants regarding the ECO dimension. On visual inspection of the ETH scores, it became clear that ETH had a very high degree of consistency. On item ETH1, 17 of the 22 experts gave the maximum rating of 7. On ETH2, 18 of 22 experts gave the maximum rating of 7. On ETH3, 16 of 22 experts gave the maximum rating of 7. On ETH4, 12 of 22 experts gave a rating of 7. Of all the dimensions, ETH had the most constant range of scores.

Consistently slightly lower reliabilities were found for SMEs' ratings when compared with those of FEs, indicating that FEs were more consistent in their scoring as a group. Across the schedule dimensions of TPM, CFOC, JCM and ECO, the results indicated a negligible difference between SMEs and FEs. In general, SMEs scored all dimensions marginally higher than did the FEs.

Thirdly, it is necessary to show that raters were viewing the same or a similar process. Stemler (2004) recommends factor loadings of .60 and greater for an indication that raters were scoring the same event. Results were slightly mixed: 45% of raters had factor loadings of greater than .60. and 36% of raters had loadings of between .40 and .60. The remaining 19% of raters had loadings ranging between .18 to .36. The results provide moderate, but not definitive, support that raters were scoring the same shared event.

Discussion

Assessment centres are a popular form of assessment, as they offer a direct, practical link to the required job, which is directly inferred from the observation of participant behaviour. Evidence suggests that the higher the degree of content validity underpinning an assessment method, the higher the probability that the assessment will have a high level of criterion-related validity (Sproule, 2009). The present study focused primarily on assessing a USA-developed VAC's content relevance and correspondence to a mid-level managerial job. No content validation study on ACs in South Africa could be found in the literature. In fact, there are few published content validation studies on ACs worldwide.

The present research contributes to AC literature and assessment methodology by demonstrating the importance and utility of content validation. It does so by providing theoretical support for the use of future content validation analysis studies. A practical contribution of this study is the possible re-introduction of experts in the assessment of content validity in cross-cultural contexts, and their explicit inclusion in psychometric assessment development. South African importers of AC content may use this study's techniques to validate content, so as to meet legal requirements and ensure domain relevance.

Summary of findings

Based on high average dimension scores and inter-rater reliability analysis, the findings of the present study indicate a high degree of agreement amongst experts on the facets of Competency Area – Job Correspondence (Critical Thinking, People Leadership, Communication, Task Process Management and Client Focus), and Job Competency – Simulation Match and Fidelity. The findings signify that, when dimensions are combined (determining overall content

relevance), the content of the VAC is relevant and applicable to the specified work domain for which it was designed. Contrastingly, factor analytic results show discrepancies in what expert raters were rating; not all raters can be considered to have been observing and scoring the same event. The data suggests that at least three experts were definitely not rating the same VAC as everyone else. However, in general there is enough support to indicate that most of the expert raters were scoring the same common event.

Roodt (2008) mentions that it is important to get the varying perspectives of different stakeholders when assessing AC content. It was expected that the two groups studied would provide different types of information regarding VAC content. The data indicated a slight but statistically significant difference between the two expert groups on the dimensions of TPM, CFOC, JCM, FDL and ECO. No statistically significant differences between groups were found for the dimensions of CT, PL, COMM, CMPLX or ETH, with the most group uniformity found with potential demographic fairness (PDF). It was originally thought that SMEs would be more consistent as a group, due to the extensive experience of industrial psychologists with psychometrics and behaviour measurement. However, when compared to FEs, the SMEs showed slight but consistently lower ICC alpha reliabilities. Similarly, Lievens (2001a, 2001b) found a difference in ratings between managers and student psychologists: the managers experienced more difficulty distinguishing between dimensions measured in an assessment exercise, but rated participants with a higher degree of accuracy.

At the core of identifying aspects in assessment that can potentially disadvantage a demographic group lies the concept of fairness in personnel decisions. The findings of the present study indicate that the VAC contains few elements that could potentially disadvantage any particular demographic group (age, gender, culture or language). The average score for PDF was 6.3 (out of a possible 7) and a very high level of unanimity between expert groups regarding this dimension was observed. This could be interpreted to mean that the VAC has very low levels of content that could potentially disadvantage a demographic group (age, gender, culture or language). The present study's results are consistent with a similar study conducted by Petrides *et al*. (2010), which found that AC exercise ratings based upon AC ratings show no ethnic or gender bias.

Managerial implications

Content validation techniques appear to be able to assess a variety of different forms of content, such as content-participant interaction, which may prove troublesome for other types of validation methods. The present study indicates that content created in the USA, or any other country for that matter, can be assessed for relevance and applicability for use in South Africa through the execution of a content validation study, using a content evaluation schedule and expert raters. The content of the VAC under inspection in the present study is deemed applicable and relevant for the selection of midlevel sales managers in the South African context.

Limitations of the study

Some of the present study's limitations are related to the design of the evaluation schedule. The dimensions of complexity and economic considerations showed the lowest levels of agreement amongst experts. These relatively low levels of agreement may indicate different fundamental assumptions by expert raters about these two dimensions, which resulted in less consistent ratings. This may indicate that the items constituting these dimensions could be further refined to elicit clearer understanding by experts. A second limitation could be that the majority of the study's research participants were English and Afrikaans speaking and, in terms of assessing potential demographic fairness, it may be beneficial to ensure a larger degree of diversity in future studies. A third limitation relates to the research design itself, in that it does not allow for the generalisability of results. An inter-rater test-retest reliability design would allow for this possibility. Lastly, whilst experts were used in this study, they are by no means infallible; it is recommended that a construct validation of the VAC be performed to further support the findings of this study.

Recommendations

The present study demonstrates the practical benefits of performing a content validation. As such, future research may make use of the procedures and approaches used in this study to demonstrate that analysis of content with the aid of experts is a valid form of assessment. This should strengthen the value of expert judgement and content validation analysis in the psychological community. Further research should solidify and sound the return of the use of content validation by the psychological fraternity, either as an appropriate alternative to the usual construct validation studies, or to increase the quality of substantive content before construct validation is conducted. Furthermore, it is recommended that future content validation studies explore the application of Rasch analysis to determine levels of rater bias and rating event and construct dimensionality.

Conclusion

The main research objective of this study was to determine whether the selected USA-developed VAC measures what it claims to measure, and to determine whether the content found within the VAC is suitable for the diverse South African context. Industrial psychologists and managers with experience in ACs were brought in as expert content evaluators. An evaluation schedule was developed, so that experts could rate the content on various applicable dimensions. Thereafter, a VAC was executed, with experts observing the process and then rating the content of the VAC on the schedule dimensions. Results indicated high average scores on dimensions, and a high degree of total agreement, with slight to moderate support for communality of rating event. In terms of overall content validity, this provides

support that the VAC does measure what it purports to, and that the content found within the VAC is suitable for use in the South African context.

Acknowledgements

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Authors' contributions

K.-P.M. (University of Johannesburg) conducted the research as part of his master's dissertation and wrote the majority of the article's content. G.R. (University of Johannesburg) contributed in the conceptualisation, planning and design of the research article.

References

Bakker, A.B., Van Emmerik, H., & Van Riet, P. (2008). How job demands, resources, and burnout predict objective performance. *Anxiety, Stress, & Coping, 21*(3), 309–324. http://dx.doi.org/10.1080/10615800801958637

Bodkin-Andres, G., O'Rourke, V., Grant, R., Denson, N., & Craven, R.G. (2010). Validating racism and cultural respect: Testing psychometric properties and educational impact of perceived discrimination and multiculturation for Indigenous and non-Indigenous students. *Educational Research and Evaluation, 16*(6), 471–493. http://dx.doi.org/10.1080/13803611.2010.550497

Bond, T.G., & Fox, C.M. (2007). *Applying the Rasch model: Fundamental measurement in the Human Sciences*. (2nd edn.). Mahwah, NJ: Lawrence Erlbaum Assoc.

Bosco, F.A., & Allen, D.G. (in press). Executive attention as a predictor of employee performance: Reconsidering the relationship between cognitive ability and adverse impact potential. *Journal of Applied Psychology.*

Boulet, J.R., Murray, D., Kras, J., Woodhouse, J., McAllister, J., & Amitai, Z. (2003). Reliability and validity of a simulation-based acute care skills assessment for medical students and residents. *Anesthesiology, 99*(6), 1270–1280. http://dx.doi.org/10.1097/00000542-200312000-00007

Brandt, A. (2005). Translation, cross-cultural adaptation, and content validation of the QUEST. *Technology and Disability, 17,* 205–216.

Brummel, B.J., Rupp, D.E., & Spain, S.M. (2009). Constructing parallel simulation exercises for assessment centers and other forms of behavioral assessment. *Personnel Psychology, 62,* 137-–170. http://dx.doi.org/10.1111/j.1744-6570.2008.01132.x

Burton, L.J., & Mazerolle, S.M. (2011). Survey instrument validity part 1: Principles of survey instrument development and validation in athletic training education research. *Athletic Training Education Journal, 6*(1), 27–35.

Byrne, B., & Van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing, 10*(2), 107–132. http://dx.doi.org/10.1080/15305051003637306

Chung, M., Chiang, I., Chou, K., Chu, H., & Chang, H. (2010). Inter-rater and intra-rater reliability of nursing process records for patients with schizophrenia. *Journal of Clinical Nursing, 19,* 3023–3030. http://dx.doi.org/10.1111/j.1365-2702.2010.03301.x

Cicchetti, D.V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*(4), 284–290.

Coetzee, M., & Schreuder, D. (2009). Using the career orientations inventory (COI) for measuring internal career orientations in the South African organisational context. *SA Journal of Industrial Psychology, 35*(1), 1–13. http://dx.doi.org/10.4102/sajip.v35i1.806

Cohen, S.L. (1980). Validity and assessment center technology: One and the same? *Human Resource Management, 19*(4), 1–11. http://dx.doi.org/10.1002/hrm.3930190402

Curtin, L.B., Finn, L.A., Czosnowski, Q.A., Whitman, C.B., & Cawley, M.J. (2011). Computer-based simulation training to improve outcomes in mannequin-based simulation exercises. *American Journal of Pharmaceutical Education, 75*(6), 1–6. http://dx.doi.org/10.5688/ajpe756113

De Bruin, G.P., & Buchner, M. (2010). Factor and item response theory analysis of the Protean and Boundaryless Career Attitude Scales. *SA Journal of Industrial Psychology, 36*(2), 1–11.

De Bruin, G.P., & Rudnick, H. (2007). Examining the cheats: The role of conscientiousness and excitement seeking in academic dishonesty. *South African Journal of Psychology*, *37*(1), 153–164. http://dx.doi.org/10.1177/008124630703700111

De Corte, W., Sackett, P., & Lievens, F. (2010). Selecting predictor subsets: Considering validity and adverse impact. *International Journal of Selection and Assessment, 18*(3), 260–270. http://dx.doi.org/10.1111/j.1468-2389.2010.00509.x

De Klerk, J.J., Boshoff, A.B., & Van Wyk, R. (2010). Measuring meaning in life in South Africa: Validation of an instrument developed in the USA. *SA Journal of Industrial Psychology, 39*(3), 314–325.

Edwards, D., & Leger, P. (1995). Psychometric properties of the right wing authoritarianism scale in black and white South African students. *International Journal of Psychology*, *30*(1), 47–68. http://dx.doi.org/10.1080/00207599508246973

Furnham, A., Jensen, T., & Crump, J. (2008). Personality, intelligence and assessment centre expert ratings. *International Journal of Selection and Assessment, 16*(4), 356–365. http://dx.doi.org/10.1111/j.1468-2389.2008.00441.x

Furtner, M.R., Rauthmann, J.F., & Sachse, P. (2011). The self-loving self-leader: An examination of the relationship between self-leadership and the dark triad. *Social Behavior and Personality, 39*(3), 369–380. http://dx.doi.org/10.2224/sbp.2011.39.3.369

Gelfand, M.J. (2000). Cross-cultural Industrial and Organisational Psychology. *Applied Psychology, 49*(1), 29–31. http://dx.doi.org/10.1111/1464-0597.00004

Golden, M. (1981). A measure of cognition within the context of assertion. *Journal of Clinical Psychology, 37*(2), 253–262. http://dx.doi.org/10.1002/1097-4679(198104)37:2%3C253::AID-JCLP2270370206%3E3.0.CO;2-Y

Gradige, D., & Jager, A. (2011). Psychometric properties of the wellness questionnaire for higher education. *South African Journal of Psychology, 41*(4), 517–527. http://dx.doi.org/10.1177/008124631104100410

Grahn, B., & Gard, G. (2008). Content and concurrent validity of the motivation for change questionnaire. *Journal of Occupational Rehabilitation, 18*(1), 68–78. http://dx.doi.org/10.1007/s10926-008-9122-7

Grant, K.L. (2009). The validation of a situational judgment test to measure leadership behavior. *Masters Theses & Specialist Projects.* (Paper 64). Retrieved May 7, 2012, from http://digitalcommons.wku.edu/theses/64

Greyling, L.A., Visser, D., & Fourie, L. (2003). Construct validity of competency dimensions in a team leader assessment centre. *SA Journal of Industrial Psychology, 29*(2), 10–19. http://dx.doi.org/10.4102/sajip.v29i2.97

Groenewald, T. (2004). A phenomenological research design illustrated. *International Journal of Qualitative Methods*, *3*(1), 1–26.

Hair, J.F., Black, W.C., Babin, B.J., & Anderson, R.E. (2010). *Multivariate data analysis: A global perspective* (7th edn.). Upper Saddle River, NJ: Pearson-Hall International. http://dx.doi.org/10.1016/j.jmva.2009.12.014

Hallgren, K.A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol, 8*(1), 23–34. PMid:22833776

Haynes, S.N., Richard, D.C.S., & Kubany, E.S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological Assessment, 7*(3), 238–247. http://dx.doi.org/10.1037/1040-3590.7.3.238

Hoeft, S., & Schuler, H. (2001). The conceptual basis of assessment centre ratings. *International Journal of Assessment Centre Ratings, 9,* 114–123. http://dx.doi.org/10.1111/1468-2389.00168

Hoffmann, B.J., & Meade, A. (2012). Alternate approaches to understanding the psychometric properties of assessment centers: An analysis of the structure and equivalence of exercise ratings. *International Journal of Selection and Assessment, 20*(1), 82–97. http://dx.doi.org/10.1111/j.1468-2389.2012.00581.x

Hoffman, B.J., Melchers, K G., Blair, C.A., Kleinmann, M., & Ladd, R.T. (2011). Exercises and dimensions are the currency of assessment centers. *Personnel Psychology, 64,* 351–195. http://dx.doi.org/10.1111/j.1744-6570.2011.01213.x

IBM SPSS 20.0 for Windows [Computer software]. Chicago, IL: SPSS Inc.

International Task Force on Assessment Center Guidelines. (2009). Guidelines and ethical considerations for assessment center operations. *International Journal of Selection and Assessment, 17*(3), 243–253. http://dx.doi.org/10.1111/j.1468-2389.2009.00467.x

Jackson, D.J.R., Stillman, J.A., & Englert, P. (2010). Task-based assessment centers: Empirical support for a systems model. *International Journal of Selection and Assessment, 18*(2), 141–154. http://dx.doi.org/10.1111/j.1468-2389.2010.00496.x

Kolk, N., Born, M., Van der Flier, H., & Olman, J.M. (2002). Assessment center procedures: Cognitive load during the observation phase. *International Journal of Selection and Assessment, 10*(4), 271–278. http://dx.doi.org/10.1111/1468-2389.00217

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B.J., Hrobjartsson, A. *et al*. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology, 64,* 96–106. http://dx.doi.org/10.1016/j.jclinepi.2010.03.002

Krause, D.E., Rossberger, R.J., Dowdeswell, K., Venter, V., & Joubert., T. (2011). Assessment Center Practices in South Africa. *International Journal of Selection and Assessment, 19*(3), 262–275. http://dx.doi.org/10.1111/j.1468-2389.2011.00555.x

Kunz, S. (2010). Psychometric properties of the EQ-5D in a study of people with mild to moderate dementia. *Quality of Life Research, 19*(3), 425–434. http://dx.doi.org/10.1007/s11136-010-9600-1

Lance, C.E., Woehr, D.J., & Meade, A.W. (2007). Case study: A Monte Carlo investigation of assessment center construct validity models. *Organizational Research Methods*, *10,* 430–448. http://dx.doi.org/10.1177/1094428106289395

Lauth, B., Magnusson, P., Ferrari, P., & Petursson, H. (2008). An Icelandic version of the Kiddie-SADS-PL: translation, cross-cultural adaptation and inter-rater reliability. *Nordic Journal of Psychiatry, 62*(5), 379–385. http://dx.doi.org/10.1080/08039480801984214

Lievens, F. (1999). Development of a simulated assessment center. *European Journal of Psychological Assessment, 15*(2), 117–126. http://dx.doi.org/10.1027//1015-5759.15.2.117

Lievens, F. (2001a). Assessor training strategies and their effects on accuracy, interrater reliability, and discriminant validity. *Journal of Applied Psychology*, *86,* 255–264. http://dx.doi.org/10.1037/0021-9010.86.2.255, PMid:11393438

Lievens, F. (2001b). Assessors and use of assessment centre dimensions: A fresh look at a troubling issue. *Journal of Organizational Behavior*, *22,* 203–221. http://dx.doi.org/10.1002/job.65

Lievens, F. (2009). Assessment centres: A tale about dimensions, exercises, and dancing bears. *European Journal of Work and Organizational Psychology, 18*(1), 102–121.

Lievens, F., Tett, R.P., & Schleicher, D.J. (2009). Assessment centers at the crossroads: Towards a reconceptualization of assessment center exercises. *Research in Personnel and Human Resources Management, 28,* 99–152. http://dx.doi.org/10.1108/S0742-7301(2009)0000028006

Lievens, F., & Thornton, G.C. III (2005). Assessment centers: recent developments in practice and research. In A. Evers, O. Smit-Voskuijl, & N. Anderson (Eds.) *Handbook of Selection* (pp. 243–264). Malden, MA: Blackwell Publishing.

Lui, J.W., Chan, F., Fried, J.H., Lin, C, Anderson, C.A., & Peterson, M. (2010). Roles and functions of counselling specialists: A multi-trait analysis. *Journal of Vocational Rehabilitation, 32,* 163–173.

Marais, C., Mostert, K., & Rothmann, S. (2009). The psychometric properties of translated versions of the Maslach burnout inventory – general survey. *SA Journal of Industrial Psychology, 35*(1), 175–182. http://dx.doi.org/10.4102/sajip.v35i1.838

McCartan, P.J., & Owen, D.W. (1990). Assessing assertive behaviour in student nurses: a comparison of assertion measures. *Journal of Advanced Nursing, 15,* 1370–1376. http://dx.doi.org/10.1111/j.1365-2648.1990.tb01778.x, PMid:2283448

McEnery, J.M., & Blanchard, P.N. (1999). Validity of multiple ratings of business student performance in a management simulation. *Human Resource Development Quarterly, 10*(2), 155–172. http://dx.doi.org/10.1002/hrdq.3920100206

Meiring, D. (2008). Assessment centres in South Africa. In S. Schlebusch, & G. Roodt (Eds.). *Assessment Centres Unlocking Potential for Growth* (pp. 21–31). Randburg, South Africa: Knowres Publishing.

Mokkink, L.B., Terwee, C.B., Gibbons, E., Stratford, P.W., Alonso, J., Patrick, D.L. *et al*. (2010). *BMC Medical Research Methodology, 10,* 1–11. http://dx.doi.org/10.1186/1471-2288-10-82

Morse, J.M. (1994). Designing funded qualitative research. In. N.K. Denzin, & Y.S. Lincoln (Eds.) (pp. 493–503). *Handbook of qualitative research*. London, UK: Sage. http://dx.doi.org/10.1177/104973239400400401

Mosdell, J., Balchin, R., & Ameen, O. (2010). Adaptation of aphasia tests for neurocognitive screening in South Africa. *South African Journal of Psychology, 40*(3), 250–261. http://dx.doi.org/10.1177/008124631004000304

Moyo, S., & Theron, C. (2011). A preliminary factor analytic investigation into the first-order factor structure of the Fifteen Factor Questionnaire Plus (15FQ+) on a sample of black South African managers. *SA Journal of Industrial Psychology, 37*(1), 1–22. http://dx.doi.org/10.4102/sajip.v37i1.934

Multon, K.D. (2010). Interater realiability. In N.J. Salkind (Ed.), *Encyclopaedia of research design, volume 1* (pp. 626–628). Thousand Oaks, CA: SAGE Publications, Inc. http://dx.doi.org/10.4135/9781412961288

Olckers, C., Buys, M.A., & Grobler, S. (2010). Confirmatory factor analysis of the Multi-dimensional Emotional Empathy Scale in the South African context. *SA Journal of Industrial Psychology, 36*(1), 1–8.

Oosthuizen, R.M., & Koortzen, P. (2009). Psychometric properties of the Experience of Work Life Circumstances Questionnaire and the Hopkins Symptom Checklist. *SA Journal of Industrial Psychology, 35*(1), 11–20. http://dx.doi.org/10.4102/sajip.v35i1.492

Perreault, W.D., & Leigh, L. (1989). Reliability of nominal data based on qualitative judgements. *Journal of Marketing Research, 26,* 135–148. http://dx.doi.org/10.2307/3172601

Petrides, K.V., Weinstein, Y., Chou, J., Furnham, A., & Swami, V. (2010). An investigation into assessment centre validity, fairness, and selection drivers. *Australian Journal of Psychology, 62*(4), 227–235. http://dx.doi.org/10.1080/00049531003667380

Polit, D.F., & Beck, C.T. (2004). Assessing data quality. In D.F. Polit, & C.T. Beck, *Nursing research: Principles and methods* (7th edn.) (pp. 413–444). Philadelphia: Lippincott Williams & Wilkins.

Polit, D.F., & Beck, C.T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29, 489–497. http://dx.doi.org/10.1002/nur.20147

Polit, D.F., Beck, C.T., & Owen, S.V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health, 30,* 459–467. http://dx.doi.org/10.1002/nur.20199

Ravens-Sieberer, U., & Bullinger, M. (1998). Assessing health-related quality of life in chronically ill children with the German KINDL: First psychometric and content analytical results. *Quality of Life Research, 7*(5), 399–407. http://dx.doi.org/10.1023/A:1008853819715, PMid:9691720

Rein, Z., Duclos, J., Perdereau, F., Curt, F., Apfel, A., Wallier, J. *et al*. (2011). Expressed emotion measure adaptation into a foreign language. *European Eating Disorders, 19,* 64–74. http://dx.doi.org/10.1002/erv.1008

Republic of South Africa. (1998). *Employment Equity Act, No. 55 of 1998*. Government Gazette 400, 19370. Cape Town: Government Printers.

Risavy, S., & Hausdorf, P.A. (2011). Personality testing in personnel selection: Adverse impact and differential hiring rates. *International Journal of Selection and Assessment, 19*(1), 18–30. http://dx.doi.org/10.1111/j.1468-2389.2011.00531.x

Roodt, G. (2008). Descriptive content analysis. In S. Schlebusch, & G. Roodt (Eds.). *Assessment Centres Unlocking Potential for Growth* (pp. 237–252). Randburg, South Africa: Knowres Publishing.

Roodt, G. (2009). Validity: Basic concepts and measures. In C.D. Foxcroft, & G. Roodt (Eds.), *Introduction to psychological assessment in the South African context* (3rd edn.) (pp. 55–64). Johannesburg, South Africa: Oxford University Press.

Roos, W., & Van Eeden, R. (2008). The relationship between employee motivation, job satisfaction and corporate culture. *SA Journal of Industrial Psychology, 34*(1), 54–63. http://dx.doi.org/10.4102/sajip.v34i1.420

Rothmann, S., Mostert, K., & Strydom, M. (2006). A psychometric evaluation of the Job Demands-Resources scale in South Africa. *SA Journal of Industrial Psychology, 32*(4), 76–86. http://dx.doi.org/10.4102/sajip.v32i4.239

Rourke, L., & Anderson, R. (2004). Validity in quantitative content analysis. *Educational Technology Research and Development, 52*(1), 5–18. http://dx.doi.org/10.1007/BF02504769

Rubio, D., Berg-Weger, M., & Tebb, S.S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Social Work Research*, *27*(2), 94–104. http://dx.doi.org/10.1093/swr/27.2.94

Russel, 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. http://dx.doi.org/10.1111/j.2044-8325.1995.tb00686.x

Sackett, P. (1987). Assessment centers and content validity: Some neglected issues. *Personnel Psychology, 40*(1), 13–25. http://dx.doi.org/10.1111/j.1744-6570.1987.tb02374.x

Schreuder, D., & Coetzee, M. (2010). An overview of industrial and organisational psychology research in South Africa: A preliminary study. *SA Journal of Industrial Psychology, 36*(1), 1–11. http://dx.doi.org/10.4102/sajip.v36i1.903

Schwartz, S.H., Melech, G., Lehmann, A., Burgess, S., Harris, M., & Owens, V. (2001). Extending the cross-cultural validity of the theory of basic human values with a different method of measurement. *Journal of Cross-Cultural Psychology, 32,* 519. http://dx.doi.org/10.1177/0022022101032005001

Shoukri, M.M., Asyali, M.H., & Donner, A. (2004). Sample size requirements for the design of reliability study: review and new results. *Statistical Methods in Medical Research, 13,* 251–271.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86(2),* 420–428. http://dx.doi.org/10.1037/0033-2909.86.2.420

Sproule, C.F. (2009). *Rationale and research evidence supporting the use of content validation in personnel assessment*. A monograph of the International Personnel Assessment Council (pp. 1–45). Retrieved from http://www.ipacweb.org

Steadman, R.H., Coates, W.C., Huang, Y.M., Matevosian, R., Larman, B.R., McCullough, L. *et al*. (2006). Simulation-based training is superior to problem-based learning for the acquisition of critical assessment and management skills. *Critical Care Medicine, 34*(1), 151–157. http://dx.doi.org/10.1097/01.CCM.0000190619.42013.94

Stemler, S.E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). Retrieved July 17, 2013 from http://PAREonline.net/getvn.asp?v=9&n=4

Storm, K., & Rothmann, S. (2003). A psychometric analysis of the Utrecht Work Engagement scale in the South African Police Service. *SA Journal of Industrial Psychology, 29*(4), 62–70. http://dx.doi.org/10.4102/sajip.v29i4.129

Swart, C., Roodt, G., & Schepers, J.M. (1999). Itemformaat, differensiële itemskeefheid en die faktorstruktuur van 'n selfvoltooiingsvraelys [Differential item functioning and the factor structure of a self-administered questionnaire]. SA *Journal of Industrial Psychology, 25*(1), 33–43.

Tett, R.P., Guterman, H.A., Bleier, A., & Murphy, J.M. (2000). Development and content validation of a 'Hyperdimensional' taxonomy of managerial competence. *Human Performance, 13*(3), 205–251. http://dx.doi.org/10.1207/S15327043HUP1303_1

Thornton, G.C. III, & Gibbons, A.M. (2009). Validity of assessment centers for personnel selection. *Human Resource Management Review, 93,* 169–187. http://dx.doi.org/10.1016/j.hrmr.2009.02.002

Tuttle, R.P., Cohen, M.H., Augustine, A.J., Novotny, D.F., Delgado, E., Dongilli, T.A. *et al*. (2007). Utilizing simulation technology for competency skills assessment and a comparison of traditional methods of training to simulation-based training. *Respiratory Care, 52*(3), 263–270. PMid:17328824

Twigge, L., Theron, C., Steel, H., & Meiring, D. (2005). A psychometric investigation into the use of an adaptation of the Ghiselli predictability index in personnel selection. *SA Journal of Industrial Psychology, 31*(1), 18–30. http://dx.doi.org/10.4102/sajip.v31i1.178

Uebersax, J.S. (1988). Validity inferences from interobserver agreement. *Psychological Bulletin*, *104,* 405–416. http://dx.doi.org/10.1037/0033-2909.104.3.405

Uebersax, J.S. (2007). *Intraclass correlation and related methods*. Retrieved November 11, 2012, from http://www.john-uebersax.com/stat/icc.htm

Uebersax, J.S., & Grove, W.M. (1993). A latent trait finite mixture model for the analysis of rating agreement. *Biometrics*, *49*, 823–835. http://dx.doi.org/10.2307/2532202

Van de Vijver, F., Tanzer, N.K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue euroéenne de psychologie appliquée, 54,* 119–135. http://dx.doi.org/10.1016/j.erap.2003.12.004

Van Vuuren, L.J. (2010). Industrial Psychology: Goodness of fit? Fit for goodness? *SA Journal of Industrial Psychology, 36*(2), 1–16. http://dx.doi.org/10.4102/sajip.v36i2.939

Visser, D., & Viviers, R. (2010). Construct equivalence of the OPQ32n for Black and White people in South Africa. *SA Journal of South Africa, 36*(1), 1–12. http://dx.doi.org/10.4102/sajip.v36i1.748

Wright, P., & Craig, M. (2011). Tool for assessing responsibility-based education (TARE): Instrument development, content validity, and inter-rater reliability. *Measurement in Physical Education and Exercise Science, 15*(3), 204–219. http://dx.doi.org/10.1080/1091367X.2011.590084

Wynd, C.A., Schmidt, B., & Schaefer, M.A. (2003). Two quantitative approaches for estimating content validity. *Western Journal of Nursing Research*, *25*, 508–518. http://dx.doi.org/10.1177/0193945903252998, PMid:12955968

Yaghmaie, F. (2003). Content validity and its estimation. *Journal of Medical Education, 3*(1), 25–27.