Australian Journal of Teacher Education

Volume 41 | Issue 7 Article 5

2016

The Development of the Stereotypical Attitudes in HPE Scale

Justen P. O'Connor
Monash University, Australia, justen.oconnor@monash.edu

Dawn Penney
Monash University, dawn.penney@monash.edu

Laura Alfrey
Monash University, laura.alfrey@monash.edu

Sivanes Phillipson
Monash University, Australia, sivanes.phillipson@monash.edu

Shane N. Phillipson
Monash University, Australia, shane.phillipson@monash.edu

Ruth Jeanes
Monash University, Australia, ruth.jeanes@monash.edu

Recommended Citation

O'Connor, J. P., Penney, D., Alfrey, L., Phillipson, S., Phillipson, S. N., & Jeanes, R. (2016). The Development of the Stereotypical Attitudes in HPE Scale. *Australian Journal of Teacher Education, 41*(7). http://dx.doi.org/10.14221/ajte.2016v41n7.5

The Development of the Stereotypical Attitudes in HPE Scale

Justen P. O'Connor
Dawn Penney
Laura Alfrey
Sivanes Phillipson
Shane N. Phillipson
Ruth Jeanes
Monash University

Abstract: This study reflects that teacher education in Health and Physical Education (HPE) has long grappled with the challenge of how to disrupt pre-service teachers' (PSTs) established attitudes about HPE that may limit their capacity to positively engage with a diverse student population. This paper describes the development, validation and interpretation of the Stereotypical Attitudes in Health and Physical Education scale (SAHPE) for use in teacher education institutions. The scale was developed as a means of exploring the extent to which PSTs perpetuate or reject discriminatory attitudes and stereotypes that have been identified as having some historical and cultural acceptance in Health and Physical Education. It was designed as a tool to generate data that can be used by teacher educators with PSTs to better understand and problematise stereotypical attitudes that ultimately impact diversity and inclusion in HPE classrooms. This paper details the design process and pilot research that enabled validation of the scale for use by Australian teacher educators. For the purposes of validation, the SAHPE scale was administered to 109 pre-service teachers at one Australian university. Rasch modelling and confirmatory factor analysis using AMOS version 20.0 was employed to determine the measurement properties of the instrument and the construct validity of factors. Results from the study showed that a 17-item factor structure of the SAHPE is a valid and reliable predictor of a construct related to stereotypical attitudes. The discussion addresses the future application of the SAHPE as an instrument for teacher educators to use in efforts to enhance inclusion in HPE.

Introduction

To the best of our knowledge, this research represents the first attempt to measure the variability in pre-service teachers' (PSTs) attitudes towards stereotypes that have historical and cultural acceptance in Health and Physical Education practice. Internationally, Health and Physical Education (HPE) and Physical Education (PE)[1] in particular continue to be identified as failing to challenge historical practices that reaffirm narrow and stereotypical

[1] In this paper we acknowledge the distinction between the learning area of HPE and the subject of PE, and variations in how these are represented in curricula internationally.

views about who 'belongs' in PE, and what knowledge, skills and understandings 'matter' (Penney & Evans, 2013). Teacher educators across Australia and internationally are confronted by research that repeatedly identifies PE as being structured and delivered in ways that establish and maintain particular types of hegemonic discourses, and privilege individuals who are white (Flintoff, 2012), masculine (Brown & Evans, 2004, Wright, 1997) and of high ability (Fitzgerald, 2009). Research reveals that those young people who cannot align with these discourses tend to disengage with HPE and experience systematic exclusion that can undermine their educational development and feelings of self-worth (Stidder & Hayes, 2013) and potentially, impact their identity and health (Evans, Rich, Davies & Allwood, 2008).

For HPE teacher education, the ongoing challenge is both clear and confronting: *(How) can the reproduction of discourses and practices that are openly exclusionary be effectively challenged?* While we have over three decades of research outlining how PE excludes along the lines of gender, ethnicity and (dis)ability (Evans & Davies, 1993; Evans & Penney, 2013; Stidder & Hayes, 2013), we know very little about the ongoing reproduction of exclusionary practices relating to choice and organisation of curriculum content, staffing arrangements, grouping strategies, methods of teaching and/or assessment. More particularly, we lack knowledge and understanding of how processes of legitimation and reproduction may be effectively disrupted by teacher education. Studies undertaken by Brown and Rich (2002) and Brown and Evans (2004) examined the perpetuation of gender stereotypes by early career teachers, providing rare insight into PSTs' attitudes, experiences and roles in the reproduction of exclusionary pedagogical practices. Other research has reaffirmed the challenges that PSTs may face in trying to (re-)negotiate (in the interests of greater equity) practices within the schools and departments in which they are placed (Sirna, Tinning & Rossi, 2008). As these and other researchers have shown, such practices include differentially positioning young people by virtue of their gender, ability, body shape, ethnicity and class (Flintoff, Fitzgerald & Scraton, 2003), often within the context of competitive and performative tasks (fitness testing; special rules for 'girls' in team sports; assessment of motor skill within a narrow band of competitive team sports) (Hay & lisahunter, 2006; Hay & Penney, 2013).

The research reported in this paper was therefore developed to open up new opportunities for teacher educators to directly engage PSTs in discussions about their own and others' attitudes towards inclusion and diversity in HPE. The paper reports the development of an instrument, the Stereotypical Attitudes in HPE (SAHPE) scale that can be utilised in initial teacher education (ITE) and professional learning contexts to generate data relating to stereotypical attitudes in PE and HPE. After providing a brief overview of the literature associated with gendered, cultural and body shape stereotypes, the authors describe the steps taken to develop and validate the scale through pilot work in one Australian teacher education institution. The final section of this paper discusses the future application of the revised instrument.

Inclusion and Stereotypes in HPE

As Evans and Davies (1993) emphasised, the meanings of equality, equity and inclusion are contested and need to be understood in relation to the specific social, political and educational contexts to which they are being applied. Following Evans and Davies (1993), our use of the term *inclusion* goes beyond technical concerns with the distribution of resources and/or opportunities amongst various groups. It focuses on an underlying concern with the impact of our actions as HPE professionals, specifically from a social justice perspective. Hence, in seeking to examine PSTs' attitudes towards inclusion and diversity,

our intention is to ascertain whether the many differences that exist amongst students in HPE classrooms are perceived "as a resource and a source of possibility, opportunity and creative change, rather than as a problem or barrier to be removed" (Evans & Davies, 1993, p.21). In particular, we recognised that stereotypical attitudes shut down opportunities to explore difference as a source of possibility. Stereotypes are pre-established uniform judgements based on false generalizations about the presumed characteristics of particular groups and their perceived members, with little sensitivity to difference and diversity (Anderson, 2006; Lucas, 2011). According to Lucas (2011) stereotyping is problematic for two reasons: 1) The generalisations on which the stereotypical judgement is based are likely to not actually be true; and 2) even if the stereotype has some merit as a generalisation, the stereotype will nearly always turn out to be false in any individual case.

Consistent exposure to stereotypes can lead to what Steele & Aronson (1995) term "stereotype threat" producing self-handicapping strategies (Stone, 2002); increased anxiety and poorer motor performance (Laurin, 2013) and a reduced likelihood that individuals will pursue a particular domain of study, such as mathematics (Good, Aronson, & Harder, 2008). Andersen (2006) challenges the idea that stereotypes are largely individualistic free-floating ideas originating in people's heads, rather, stereotypes are "…deeply embedded in the structure of institutions" (p.84). Other research has revealed PE teachers' philosophies as heavily entangled in sport and performance discourses (Alfrey, Cale & Webb, 2012; Green, 2003) that reflect the position of sport in wider society and promote the superiority of hegemonic masculinity, high ability and whiteness (Brown & Evans, 2004, Flintoff, 2012).

From our perspective as teacher educators, knowing more about how stereotypes are rejected or perpetuated in PE and HPE systems, can directly assist our endeavours to promote greater understanding of inclusion and diversity and more specifically, open up spaces to engage teacher educators with alternative discourses of difference. In our view ITE represents a key context within which PSTs' philosophies can be disrupted, a critical pedagogical lens can be fostered, and inclusive practices developed. Indeed, we believe ITE must provide safe opportunities for PSTs to challenge ongoing inequities in PE and HPE more generally. This research provides one way to apply "…a critical appraisal of the value assumptions about ability, 'race' and gender differences" that are reportedly held within the institution (Evans & Davies, 1993, p.21).

Developing the SAHPE: Insights from Literature

A literature review informed the scale development and more specifically, identified the stereotypical attitudes to be reflected in items (see Methodology for further details of the search processes informing the review). While research has explored HPE teachers' and teacher education students' knowledge and beliefs (Ennis, 1994), and more particularly, pursued the value orientations that shape interpretation and enactment of curriculum (Ennis, 1992; Ennis & Chen, 1995; Gillespie, 2011), we sought to more directly examine stereotypical attitudes relevant to the HPE learning area in Australia connecting with different forms of exclusion and inclusion. This is a notably narrower curriculum context than PE curricula internationally, but features issues relating to inclusion and diversity that will be recognisable to professionals in other countries. Below we outline how gender, culture, the body and ability, formed the focus for SAHPE. We acknowledge that we have been selective in our coverage of the range of stereotypes that are perpetuated in PE and HPE and recognise the potential to extend this work to address other stereotypes (relating, for example, to sexuality and social class). Whilst we discuss each as discrete foci, we acknowledge that stereotypes are intersectional and form interlocking systems of domination (Andersen, 2006). The intent was to develop a pool of items representing different

stereotypical attitudes that point to an overall latent variable (stereotypical attitude) for further analysis as observed data.

In relation to gender, Green, Smith, Thurston and Lamb (2007) highlight that gendered patterns of provision that lie in the hidden curriculum of competitive team games and school sporting teams remain normalised amidst numerous curriculum reforms in the UK. Other research (Cassidy et al., 2008; Wilkinson, Littlefair & Barlow-Meade, 2013) has indicated that teachers' perceptions of male and female students' abilities continue to be informed by gendered assumptions. In the context of senior PE in Queensland, Hay and Macdonald (2010) highlighted the consensus amongst teachers was that males excel in terms of physical ability whilst females are more academically able. These assumptions have consequences for the ways in which males and females view and act within their social worlds (Wright & Burrows, 2006), and repercussions for the ways in which teachers identify ability and achievement (Hay & Macdonald, 2010).

Similarly, research points to assumptions about culture and ethnicity as impacting student experiences and achievements in HPE. Fitzpatrick's (2011) work with Maori and Pacifica youth further highlights the need for enhanced understandings of HPE in particular cultural contexts, and how as a specialised field of knowledge it expresses and legitimates dominant culture and class hierarchies (Bernstein, 1990; 2000). Wright and Burrows (2006) point us towards work by a range of authors who have provided valuable insights into the ways in which "physical 'ability' has been used to differentiate between different races and ethnicities" (p.286). They discuss the false dichotomisation of 'black bodies' (they include Maori, Pasifica, Australian Aboriginal and Afro-American bodies) and 'white bodies' with the former assumed to have physical ability, and the latter assumed to have intellectual ability. The inherent limitations of such conceptions are clear, as is the need in increasingly diverse classrooms for HPE professionals to move away from such homogenisation.

In HPE the appearance of the body is fundamental to teachers' judgements of 'ability' and students' perceptions about themselves and their peers in HPE. Research (Evans, 2004; Hay & lisahunter 2006; Hay & Macdonald, 2010; Penney & lisahunter, 2006) reaffirms the need to critically examine how teachers and students 'read' abilities onto bodies in HPE and has shown that attitudes towards 'the body' are central to the marginalisation or exclusion of students. PE has long been identified as a curriculum subject where the physicality of students is continually and publicly exposed to be judged by others (Fitzgerald & Stride, 2012). Furthermore, HPE is a context that sees 'the body' brought to the fore in students' developing understandings about health and about who is healthy (see for example, Burrows & Wright, 2006, 2010). PE arguably promotes a 'paradigm of normativity' (Fitzgerald, 2005: 41) that privileges particular types of bodies, abilities and embodied competencies including aggression, competition and masculinity (Bramham, 2003). Fitzgerald and colleagues' (Fitzgerald, 2005; Fitzgerald & Stride, 2012) research has specifically brought to the fore the experiences of students with disabilities, revealing that both teachers and classmates viewed young people with disabilities as incompetent and unable to fully benefit from PE classes. Fitzgerald (2005) particularly highlighted that teachers draw on normative discourses of 'preserving' and 'protecting' to justify the exclusion of young people with disabilities from certain activities.

In relation to the various equity issues discussed here, research has rendered the HPE profession as steadfastly resistant to change, with individualistic and instrumental understandings of health and the body reigning supreme (Alfrey & Gard, 2014; Garrett & Wrench, 2008; Pronger, 1995; Skelton, 1993; Wright, 2000). This research sought to challenge this resistance through the development of a scale that could be used to understand the extent to which stereotypical attitudes are rejected or perpetuated in HPE, and with those understandings, support future teachers to consider their own preconceived judgements.

Methodology: SAHPE Development Process

In developing an instrument in the social sciences, there is a need to confine it to a manageable number of items for respondents. We could not attempt to address all elements of stereotypical attitudes that may feature in HPE and acknowledge that some people will identify gaps in the issues that we sought to incorporate into the scale. Having determined that the instrument was to measure a range of stereotypical attitudes in physical education and HPE and be utilised in teacher education and professional learning contexts, a search for existing instruments was conducted. A defined set of search terms (including inclusive; inclusion; exclusive; exclusion; equity; stereotype; gender; culture; race; class; body; teaching; practice; pedagogy; physical education; organisations; workplace; instrument; measuring; measure; scale; test; inventory) was entered into multiple databases, including ERIC, Proquest, Psychinfo and Sportdiscus. From this search a number of survey instruments were identified that addressed inclusion (Ahmmed, 2013; Brandes, McWhirter, Haring, Crowson, & Millsap, 2012; Costello & Boyle, 2013; Ernst & Rogers, 2009; Jeong & Block, 2011; Sharma, Loreman, & Forlin, 2012; Wilczenski, 1995) but these were not designed specifically for a HPE context and or focused primarily on one area (i.e. disability). As such these scales were deemed unsuitable by the researchers for exploring stereotypical attitudes amongst HPE PSTs. The literature was consequently utilised to generate an item pool (DeVilles, 2003). A total of 31 research papers (seven not related to HPE), largely qualitative in nature, were tabulated and scanned for content related to stereotypical attitudes in HPE.

The intent in then developing a pool of items was to focus on particular stereotypical attitudes that point to an overall latent variable (stereotypical attitude). An initial item pool comprised 42 items. These were cut on the basis of a priori criteria including lack of clarity, questionable relevance and undesirable similarity to other items. Multiple negative, double-barrelled and reading difficulty level items were also removed leaving an item pool of 20 items for formatting (DeVilles, 2003).

In order to establish content validity, the items at this point were sent for external review. Three international reviewers[2] were provided with a working definition of the construct and asked to rate each item with respect to its relevance and wording clarity in accordance with steps outlined by DeVilles (2003). In addition, six local HPE experts reviewed and rated items in a workshop. Taking into account recommendations from the nine reviewers, 12 of the 20 items were subject to significant changes to wording. With this improvement, all 20 items were deemed relevant enough to be administered to a pilot sample.

The SAHPE scale comprised of 20 items that asked teachers to consider various statements related to gender, cultural, disability or body shape stereotypes identified within the literature as existing in HPE (Table 1). The statements in the SAHPE are representative of stereotypical attitudes for which PSTs were asked to rate their agreement with each statement in the item on a four-point scale ranging from 1 (entirely agree) to 4 (entirely disagree). For example, Item 4 is "South East Asian students are more suited to racket sports." The responses to items 15, 16 and 19 were reverse coded.

[2] Seven reviewers were initially invited to participate.

How much do you agree or disagree with the following statements?	
#	*Question*
1	Girls learn best in aerobics and dance activities whilst boys learn best in activities like football.
2	Tall people move and think slower when participating in sports.
3	Girls are more masculine (man-like) when playing sport[a].
4	South East Asian students are more suited to racket sports.
5	It is important to speak slowly and clearly when working with disabled students.
6	Successful HPE students need to be highly physically skilled.
7	In HPE, boys respond to direct orders whereas girls need more 'mothering.'
8	Aboriginal and Torres Strait Islanders excel in ball sports.
9	Disabled students have a greater risk of injury in HPE.
10	Girls prefer to tone up in the gym whereas boys prefer to develop muscle and strength.
11	Aboriginal and Torres Straits Islanders want to play indigenous games in HPE to highlight how different their cultures and people are[b].
12	Female students are less likely to be able to answer complex questions related to tactics in HPE class.
13	Pacific islanders make great rugby players.
14	Boys do not enjoy dance in HPE[c].
15	Overweight people can always successfully complete HPE studies[#]*.
16	In HPE, girls possess the same will to compete as the boys[#]*.
17	Communication with disabled people in sport is best conducted through their carer.
18	Racial differences can lead to advantages in certain sports.
19	Girls who play rugby are just as feminine as girls who enjoy dance[#].
20	An 'overweight' HPE teacher can be an excellent role model for students*

Table 1: SAHPE questions

[#] Items reversed coded

** Items removed from final SAHPE*

a. Item wording changed from: *Aggressive girls are more…*

b. Item wording changed from: *Teaching Aboriginal and Torres Straits Islander games in HPE is important to…*

c. Item wording changed from: *Boys who enjoy dance in HPE generally don't fit in.*

Pilot Participants

A total of 109 first year students enrolled in a Bachelor of Education degree, specialising in HPE were invited to complete the survey during the first week of their degree studies. Following data screening, 108 respondents were included for analysis. Table 2 provides a summary of participant characteristics.

Procedure

Following ethical clearance, participants were invited to undertake the SAHPE. The invitation was made during the first week of participants' degree studies, by an independent research assistant not known to the participants. A room and digital equipment was provided where participants could voluntarily complete the SAHPE and additional questions relating to inclusion using Qualtrics, an online survey platform. Participants were encouraged to answer each question to the best of their ability by the independent administrator and were limited to one attempt. The total survey time, encompassing the SAHPE subscale took between 15 and 28 minutes for the majority (82%) of the participants.

Characteristics	No.	Percentage
Age		
18-21 years	95	87%
22-25 years	10	9%
26+ years	4	4%
Gender		
Female	64	59%
Male	45	41%
Schooling		
Govt	54	50%
Independent	23	21%
Catholic	32	29%
Geographical identity		
Australian	103	94%
European	4	4%
South-East Asian	2	2%

Table 2: Subject Demographics

Whilst completing the pilot survey, participants were prompted in the survey to comment on question clarity. Analysis of comments and discussion with responders within one week of piloting the test identified concerns with wording on Items 3, 11, 14 and 16. Following the validation (outlined in more detail in the following section), items 15, 16 and 20 were removed and minor wording changes to the remaining items were made to enhance clarity (Table 1).

Approach to Validation

Two types of main analysis were used to validate the 20-item SAHPE, including Rasch modelling and confirmatory factory analysis (CFA). In order to confirm the measurement structure of the SAHPE, Rasch modelling using WINSTEP version 3.74 was employed, providing information on item functioning so as to determine the suitability of each item in measuring the theoretical constructs of SAHPE, and whether or not the four-point rating scale is being used by respondents as intended. Rasch modelling determined if the instrument was able to distinguish between persons that differ in their attitudes and stereotypes toward gender, culture, ability, and the body. It is important to note that the Rasch model produces estimates of item difficulty and person ability using the logit measurement scale. It is therefore possible to directly compare item difficulty with person "ability".

Linacre's (2004) eight guidelines for optimising rating scales were used as the benchmark by which the categories in SAHPE were evaluated for their performance. The guidelines serve as useful criteria to determine the underlying measurement structure of the scale matches the theoretical assumptions of SAHPE. The guidelines are:

1. There should at least be 10 counts of responses for each of the four categories, from entirely agree through to entirely disagree.
2. There needs to be a uniform distribution of observations across the categories - peaks for frequently used categories and flatter long tails for less frequently used categories.
3. Average measures from Category 1 to 4 need to progress monotonically.
4. An outfit mean-square of less than 2.0 is preferred for well performing categories because higher levels than 2.0 indicates unexplained randomness that poses a threat to the objectivity of the measurement scale.

5. There needs to be an orderly series of step calibrations that advance in a monotonic manner.
6. For guidelines 6, 7 and 8, Linacre (2004) recommends that the distances or gaps between categories should be larger than 1.4 logits and less than 5.0 logits. An examination of the probability curve gives a diagnostic view of the gaps – each category should have individual peaks cutting across each other at the step calibration or threshold measure. A faulty category would appear as a flat line, and intersections between categories would have little gap or too big a gap.

The interpretations of the fit statistics are based on Bond and Fox (2015), Linacre (2002b; 2015), Wright and Stone (1999) and Scanlan, Lannin, and Hoffmann (2015). Item and person functioning is examined using infit (information-weighted) and outfit Mean Square (MnSq) statistics, with values of 1.0, representing "perfect" fit with the Rasch model. However, acceptable MnSq values are .8 – 1.2 ("excellent" fit), .5 -1.5 ("acceptable" fit). Items that are outside these ranges are termed misfitting, meaning that the item or person is mis-measured and could be removed from the measurement model.

Infit and outfit statistics can be transformed into approximately normalized *t* distributions with infinite degrees of freedom (or Z distribution), commonly referred to as infit Zstd and outfit Zstd (Bond & Fox, 2015). When the responses fit the Rasch model, the expected values of Zstd are 0±2. Last, point-measure correlations should be positive in value if the item contributes to the measurement scale as expected.

Winsteps produces indices of reliability, including *item separation reliability*, *item separation index* and *item strata index*. Item separation reliability indicates the reliability in the item measures, interpreted similarly to Cronbach's alpha with values of 1.0 indicating perfect reliability. As mentioned earlier, items should differ in terms of their perceived difficulty if the measurement scale is considered to be adequate for its purpose. An adequate item separation index would be >2.0, indicating at least two levels of difficulty. The item strata index indicates the number of distinct groups of items.

In terms of a 4-point Likert scale, rating scale functioning was examined using the guidelines suggested by (Linacre, 2002a), including the outfit MnSq should be less than 2.0 and Rasch-Andrich thresholds advance by less than 5.0 logits. In addition, the Rasch-Andrich thresholds should advance monotonically and, finally, the average measures should advance monotonically. Well-functioning category response curves are uniformly distributed and peak towards 1.0, indicating monotonic step progressions and consistent category diversity in item responses and person ability (Linacre, 2004).

Importantly, Rasch modeling assumes a certain level of uni-dimensionality in its measurement due to its ability to process cumulative measures and 'compare' persons and/or items (Boone, Staver, & Yale, 2014) through the provision of acceptable reliability estimates. Additionally, when item properties impact response patterns in a unified manner while maintaining loyalty to the theoretical construct, uni-dimensionality can be achieved (Edelen & Reeve, 2007). Acceptable ranges of MnSq infit and outfit are 0.9 – 1.1 and 0.7 – 1.3 while person separation indices above 0.9 are also considered indicators of uni-dimensionality (Tennant & Pallant, 2006).

To confirm the factor structure of the SAHPE, CFA was employed with a maximum likelihood estimation using AMOS version 20.0. In order to maintain a good balance of items for each of the different conceptual themes within stereotypical domain, covariances of error measures were added as recommended by Byrne (2010). A number of goodness of fits, including traditional and non-traditional fit values, was reported in this study as evidence of SEM model fits. A non-significant χ^2 ($p > .001$) is preferred as it shows a small discrepancy between the hypothesized model and the population. Ratio values of less than 2.00 in a χ^2/df indicate a good-fitting model (Bollen, 1989). Other goodness of fits that are reported in this

study include the Goodness of Fit Index (GFI) (Joreskog, 1993), Comparative Fit Index (CFI) (Bentler, 1990) (and the Root Mean Square Error of Approximation (RMSEA) (Bollen, 1989) and the PCLOSE. These indexes were chosen because they provide stringent measures of fit in consideration of sample variances and have been frequently quoted as the sufficient indicators of fit along with the traditional chi-squares (Bollen & Long, 1993; Hu & Bentler, 1999).

Results

Three types of results are presented. Firstly, descriptive statistics outlining mean, standard deviations of the raw scores along with overall and item deleted Cronbach Alphas. This is followed by item fit statistics of the Rasch rating scale model including item and person measures and reliabilities, rating item and person fits and point-measure correlations of the SAHPE. Finally, the model fits and factorial validity of the SAHPE are presented to confirm the one-factor CFA model with an elaboration of the best fitting models between a 20-item construct and 15-item construct. Both traditional (χ^2/df) and alternative fit indices (GFI, CFI RMSEA and PCLOSE) are reported for the model fits.

An initial overview of the data showed one participant with no response at all for the 20 items. This participant was removed listwise from the data. There were 108 valid responses out of the 109 responses with the missing data found to be at random. These missing data were replaced with series means. Table 3 shows the mean and standard deviation of the 20 items of SAHPE. The overall Cronbach Alpha for 108 complete responses to the 20 items SAHPE was .86. Three of the items, if deleted, would result in a slightly higher reliability coefficient of .87, and notably, these items were: Item 15; Item 16; and Item 20.

Rasch Modelling

Rasch modeling of the responses to the 20-item scale result showed the mean person ability was .52, indicating that the sample of 108 student teachers found the items to be a mix of easy and hard items. The person fit measure, with an infit MnSq of 1.02 and outfit MnSq of 1.00, were close to 1.00 as expected for a well-fitting model (Table 4). The *t* values of -.20 for both infit and outfit were also indicative of acceptable fit, as was the person separation reliability of 0.81.

No.	Items	Mean	SD	Cronbach's Alpha if Item Deleted
1	Gendered Stereotypes	2.84	0.99	.85
2	Body Stereotypes	3.29	0.84	.85
3	Gendered Stereotypes	2.90	0.91	.84
4	Cultural Stereotypes	3.16	0.81	.84
5	Ability Stereotypes	2.07	0.79	.86
6	Body Stereotypes	2.64	0.82	.85
7	Gendered Stereotypes	3.10	0.75	.85
8	Cultural Stereotypes	2.87	0.75	.85
9	Ability Stereotypes	2.55	0.80	.85
10	Gendered Stereotypes	2.34	0.90	.85
11	Cultural Stereotypes	2.67	0.84	.85
12	Gendered Stereotypes	3.33	0.77	.85
13	Cultural Stereotypes	2.56	0.86	.85
14	Gendered Stereotypes	2.97	0.88	.85
15	Body Stereotypes	3.17	0.82	.87
16	Gendered Stereotypes	2.98	0.88	.87
17	Disability Stereotypes	2.65	0.76	.85
18	Cultural Stereotypes	2.77	0.84	.85
19	Gendered Stereotypes	2.85	0.87	.86
20	Body Stereotypes	2.45	0.83	.87

Table 3: Mean, standard deviations and Cronbach Alpha for 20-item SAHPE (N=108)

Measure summary		Item	Person
	Mean (SD adjusted)	.00 (.71)	.52 (.79)
	Reliability of estimate	.96	.81
Fit statistics			
Infit Mean Square			
	Mean (SD)	1.00 (.38)	1.02 (.54)
	Infit *t* (SD)	-.40 (2.80)	-.20 (1.90)
Outfit Mean Square			
	Mean (SD)	1.00 (.41)	1.00 (.54)
	Outfit *t* (SD)	-.30 (2.80)	-.20 (1.90)

Table 4: Summary of item and person estimates for SAHPE 20-item scale (N = 108)

Importantly, the item separation reliability was .96, reflecting a good spread of items on the measurement scale (Bond & Fox, 2015). The overall item fit measures of 1.00 for both infit MnSq and outfit MnSq are as expected for a well-fitting Rasch model. The infit and outfit *t* values were also close to .00, also indicative of a well-fitting model. However, the standard deviations of the infit and outfit *t* values were above 2.0, indicating some faulty item functioning, which warranted further examination of the item structures. A close look at the item structures of the 20 items showed that a number of items had large infit and outfit t values (>2.0). These values are bolded in Table 5.

No.	Items	Measure	Error	IN.MSQ	IN.ZSTD	OUT.MSQ	OUT.ZSTD	PTME[a]
1	Gendered Stereotypes1	-0.06	0.13	1.26	2.02	1.24	1.77	0.56
2	Body Stereotypes1	-0.95	0.14	0.88	-0.92	0.80	-1.40	0.65
3	Gendered Stereotypes2	-0.17	0.13	0.81	-1.63	0.78	-1.84	0.71
4	**Cultural Stereotypes1**	-0.67	0.14	0.67	**-3.13**	0.64	**-2.95**	0.71
5	Ability Stereotypes1	1.47	0.14	1.18	1.29	1.11	0.80	0.40
6	Body Stereotypes2	0.32	0.13	0.95	-0.35	0.98	-0.14	0.50
7	**Gendered Stereotypes3**	-0.56	0.14	0.63	**-3.52**	0.62	**-3.23**	0.65
8	**Cultural Stereotypes2**	-0.11	0.13	0.72	**-2.61**	0.74	**-2.16**	0.57
9	**Ability Stereotypes2**	0.50	0.13	0.72	**-2.47**	0.72	**-2.40**	0.63
10	Gendered Stereotypes4	0.91	0.14	0.94	-0.39	0.91	-0.70	0.65
11	Cultural Stereotypes 3	0.27	0.13	0.99	-0.06	0.98	-0.10	0.51
12	Gendered Stereotypes5	-1.05	0.15	0.79	-1.76	0.71	-2.02	0.63
13	Cultural Stereotypes4	0.46	0.13	0.97	-0.22	0.93	-0.53	0.57
14	Gendered Stereotypes6	-0.31	0.13	1.02	0.19	1.03	0.29	0.54
15*	**Body Stereotypes3**	-0.69	0.14	1.44	**3.24**	1.65	**3.91**	**0.23**
16*	**Gendered Stereotypes7**	-0.33	0.13	1.63	**4.54**	1.92	**5.62**	**0.17**
17	Ability Stereotypes3	0.30	0.13	0.80	-1.72	0.84	-1.30	0.52
18	Cultural Stereotypes5	0.08	0.13	0.88	-1.01	1.06	0.53	0.55
19	Gendered Stereotypes8	-0.08	0.13	1.21	1.67	1.23	1.70	0.41
20*	**Body Stereotypes4**	0.67	0.13	1.48	**3.34**	1.49	**3.33**	**0.21**

Table 5: Item structure of 20-item SAHPE

[a] Point measure estimate/correlations

*These items show both larger infit and outfit *t* values and smaller point measure correlations.

Bolded items have standardised infit and outfit mean square larger than 2.0

A sequential elimination of the items with large infit and outfit *t* values were performed for the further Rasch analysis, beginning with Item 16 as it had the largest infit and outfit *t* values and also poor point measure correlation (-.17). Each time an item was removed, the remaining items' infit and outfit values were examined to ensure that the item to be removed still had poor fit structures. The next item removed was Item 20, followed by Item 15.

The removal of severely non-functioning items resulted in a 17-item scale. A close look at the 17 items functioning showed a few items (Items 1, 4, 5, 7 and 19) larger than 2.0 infit and outfit *t* values, but all the point measure correlations of these items were larger than .45, indicating that they do belong in the scale. Furthermore, removing any of these items caused the scale to worsen and other items to mis-function even more. Therefore, a 17-item scale was considered a valid and reliable construct of SAHPE. The Rasch fit indices of the 17-item scale is reported in Table 6.

Measure summary		Item	Person
	Mean (SD adjusted)	.00 (.71)	.77 (1.22)
	Reliability of estimate	.96	.86
Fit statistics			
Infit Mean Square			
	Mean (SD)	1.00 (.23)	1.02 (.54)
	Infit *t* (SD)	-.10 (1.80)	-.10 (1.70)
Outfit Mean Square			
	Mean (SD)	.99 (.25)	.99 (.52)
	Outfit *t* (SD)	-.10 (1.70)	-.20 (1.60)

Table 6: Summary of item and person estimates for SAHPE 17-item scale (N = 108)

To lend further confidence to the 17-item scale of SAHPE, the category structures of the 17-items were checked against Linacre's (2004) eight guidelines. Table 7 shows the category structure of responses in the scale, evaluated using the eight guidelines. All but Guidelines 2 and 6 are fulfilled. Guideline 2 requires a uniform distribution of observations across the categories. Initial examination of the category probability curves showed a flatter peak for category 3 and longer tails for categories 1 and 4, compared to category 2 (see Figure 1). Also, there is a lack of coherence between ratings to measure for category 1 (*Strongly Disagree*) where only 10% of those participants who chose category 1 are placed by the measure in the same category. There is also lack of coherence between measures to ratings for category 4 (*Strongly Agree*) where only 44% of measure explained participants' choice of the ratings in the same category (Table 7).

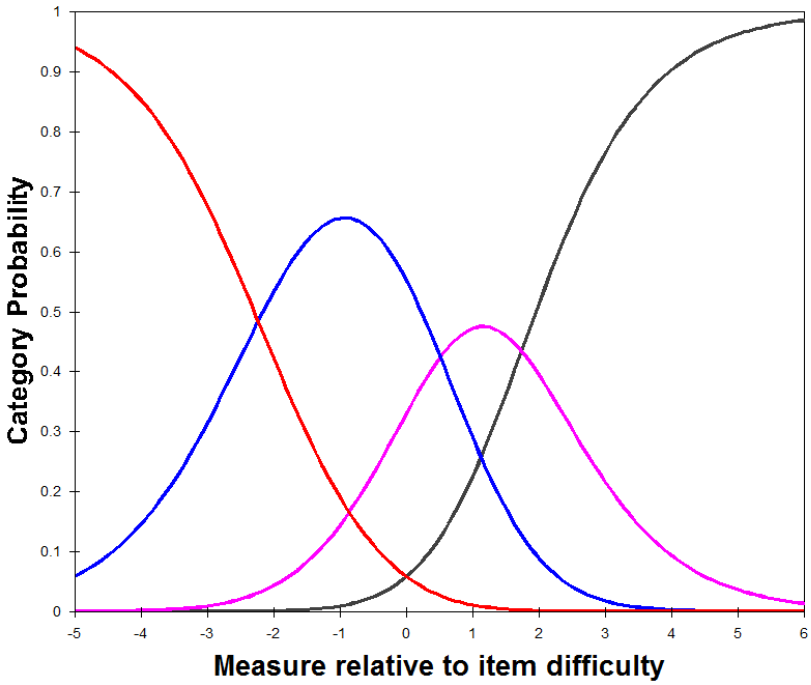

Figure 1: Probability of curve of the 4-point categories of 17-item SAHPE, showing the second category (Disagree) peaking above all the other categories

No.	Guideline	Stereotypes Attitude in Health and Physical Education Scale (17 items)[a]
1	At least 10 observation of each category	Category 1 = 107 (6%) Category 2 = 637 (34%) Category 3 = 631 (34%) Category 4 = 478 (5%)
2	Regular observation distribution	Observation distribution triangular peaked at category 2
3	Average measures advance monotonically with category	Category 1 = -.69 Category 2 = .01 Category 3 = .86 Category 4 = 2.13
4	OUTFIT mean-square less than 2.0	Category 1 = 1.28 Category 2 = 1.01 Category 3 = .85 Category 4 = .91
5	Step calibrations advance	Category 1 = None Category 2 = -2.24 Category 3= .51 Category 4 = 1.73
6	Ratings imply measures, and measures imply ratings	Coherence M->C C->M[b] Category 1 = 55% 10% Category 2 = 63% 57% Category 3 = 46% 73% Category 4 = 81% 44%
7&8	Step difficulties advance by at least 1.4 logits and by less than 5.0 logits	Category 1 = -3.39 Category 2 = -.93 Category 3 = 1.15 Category 4 = 3.01

Table 7: Rating scale category effectiveness for 4-point response category in the SAHPE using guidelines set out by Linacre (2004)

[a]Categories: 1 = Strongly Disagree, 2 = Disagree, 3 = Agree, 4 = Strongly Agree.
[b]M->C Does measure imply category? C->M Does category imply measure?

Concerns with distribution and coherence are considered minor in relation to how the categories were meeting the other guidelines in totality, especially in the sense of orderly structure and sufficient gaps between the responses to the categories (Bond & Fox, 2015). Hence, it was found that Rasch rating scale shows a 17-item valid structure for SAHPE. In order to affirm this finding, CFA was further employed to determine the best construct of the SAHPE.

Confirmatory Factor Analysis

The raw scores were used to build a full 20-item CFA to confirm the findings of the Rasch analysis. This model was poor fitting with a significant chi-square and very poor PCLOSE, and GFI of .820 and CFI of .882 (see Table 8). The model had three items with zero loading, and they were Item 15-Physical Stereotypes3, Item 16-Gendered Stereotypes7 and Item 20. Two other items, Item 19 (.07) and Item 5 had (.09) multiple squared correlations.

A second model without the three items with zero loading was built. Modification index showed a number of missing covariances and these were added to the model. This 17-item model was a good fitting model with $\chi^2/df = .974$ $p = .557$, well within the acceptable value of less than 2.00. The GFI was .901 and the CFI was 1.00, showing excellent fit. The RMSEA was also excellent with a value of .00 with a PCLOSE of .968. The lowest loading items are the Item5 and Item19 with .08 multiple squared correlations, however, the two items had significant path indicators to the SAHPE factor.

SAHPE	Chi-square (χ^2)	Degrees of freedom (*df*)	χ^2/df	GFI	CFI	RMSEA	PCLOSE
20-items	241.75	165	1.465	.820	.882	.066	.081
17-items	105.24	108	.974	.901	1.000	.000	.968
15-items	70.73	82	.86	.93	1.00	.000	.991

Table 8: The SAHPE 20-item, 17-item, and 15-item measurement model fits (N = 108)

A third model without the latter two items was built further to determine whether model fit differed significantly between a 17-item SAHPE or a 15-item SAHPE. The 15-item model had slightly better model fit (refer to Table 5) than the 17-item but the two models were not significantly different ($\Delta \chi^2 = .10$). Thus, it appears that the 17-item structure is probably the best scale for SAHPE confirming the Rasch results as well (Figure 2).

To develop the SAHPE, Rasch modelling of the responses to the full 20 items initially showed that a number of items were misfitting. However, sequential elimination of these items showed that the best fit of the responses to the model occurred when only Items 15, 16 and 20 were removed. Furthermore, the modelling showed that the 4-point Likert scale was adequate for the purpose of distinguishing between persons of differing stereotypes. Confirmatory factor analysis confirmed that the SAHPE has adequate construct validity. *Thus, the SAHPE is an instrument that can measure gendered, cultural, and ability stereotypes in students undertaking initial teacher education.* The misfitting items relating to body shape should be examined more closely to determine the reasons for their misfit, including whether or not the wording could be changed to ensure that the item is comprehensible for the respondee. Further development of the SAHPE will endeavour to address this issue and also extend the scale to addressing other stereotypical attitudes relating to, for example, social class, that were not incorporated in this initial development phase. It is anticipated that ongoing refinement of the SAHPE will lead to an improved measures of stereotypical attitudes.

Conclusion

The SAHPE is an easily administered online scale developed to better understand how PSTs either accept or reject commonly reported stereotypes within HPE, to provide data to guide teaching and learning about stereotypical attitudes and, to ascertain how stereotypical attitudes might change over time within ITE. Thus far, our results show that the 17-item structure of the SAHPE is a valid and reliable predictor of a construct related to stereotypical attitudes. Data generated by the SAHPE scale should enable teacher educators and PSTs to establish a conversation about stereotypes within ITE cohorts, and follow how PSTs' attitudes change across an ITE course. By offering a way to critically appraise the value assumptions reportedly held within the institution, the SAHPE can be utilised as an educational resource to reinforce inclusive understandings that accept young people as complex, diverse and unique individuals. The pilot work that has informed development of the scale has provided us with opportunities to engage in new, and in our view, educationally important conversations with PSTs in HPE. Research using the SAHPE with this cohort of PSTs and with others in other teacher education institutions in Australia is ongoing. Subsequent publications will shed further light on the value assumptions related to stereotyping that are reportedly held within HPE and ITE.

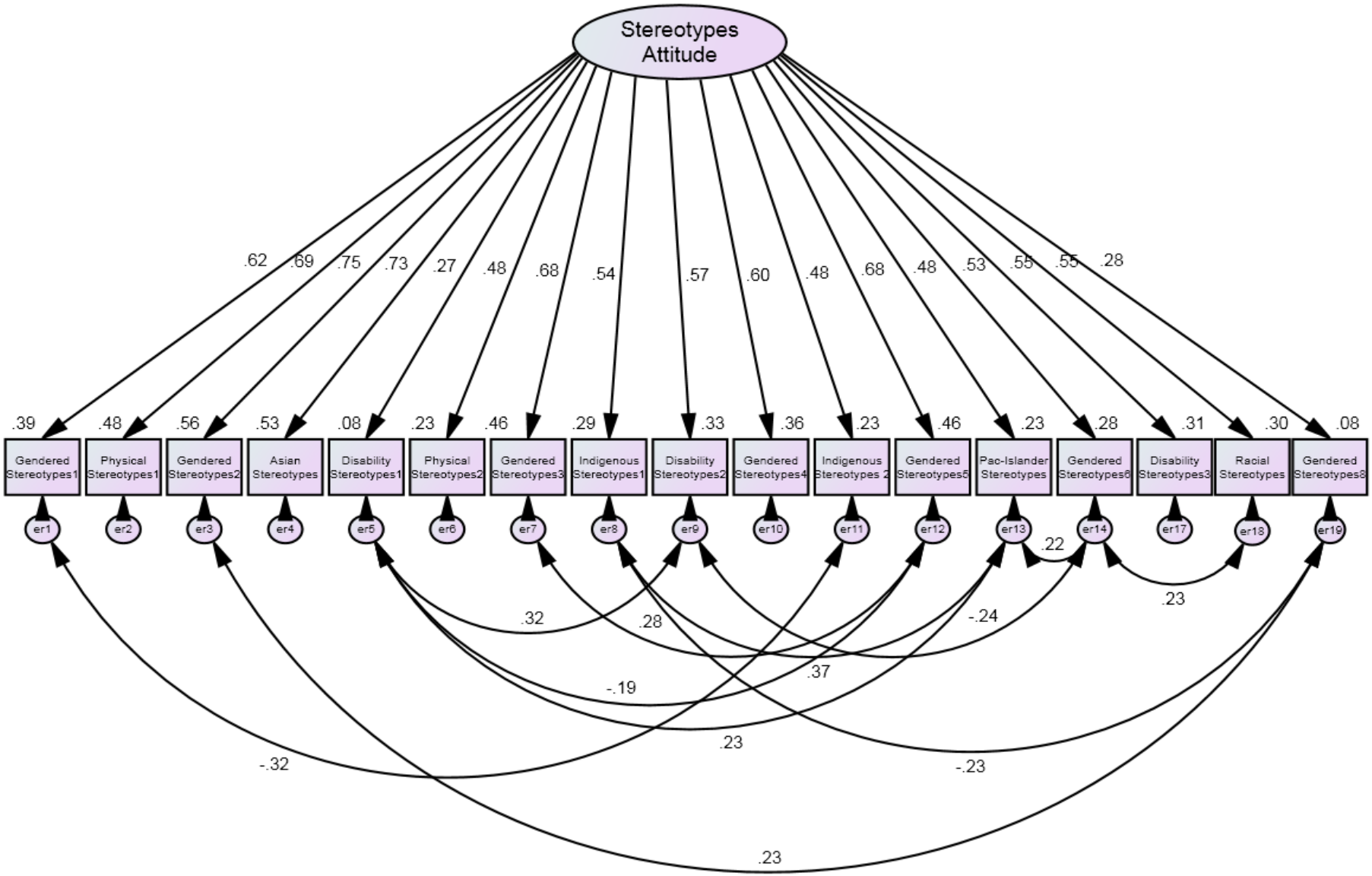

Figure 2: The 17-item measurement model of SAHPE is a good fitting model, with significant loadings and an unstandardised effect size of .34. The model shows standardised estimates and multiple squared correlations.

References

Ahmmed, M. (2013). Measuring perceived school support for inclusive education in Bangladesh: the development of a context-specific scale. *Asia Pacific Education Review*, 14(3), 337-344. http://dx.doi.org/10.1007/s12564-013-9263-z

Andersen, M. L. (2006). Race, Gender, and Class Stereotypes: New Perspectives on Ideology and Inequality. *Norteamérica, 1*(1), 69-91.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246. http://dx.doi.org/10.1037/0033-2909.107.2.238

Bernstein, B. (1990). *The structuring of pedagogic discourse. Volume IV Class, codes and control*. London: Routledge. http://dx.doi.org/10.4324/9780203011263

Bernstein, B. (2000). *Pedagogy, symbolic control and identity. Theory, research, critique* (Revised ed.). Oxford: Rowman & Littlefield.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons Inc. http://dx.doi.org/10.1002/9781118619179

Bollen, K. A., & Long, S. J. (1993). Introduction. In K. A. Bollen & S. J. Long (Eds.), *Testing structural equation models* (pp. 1-9). Newbury Park: SAGE Publications. http://dx.doi.org/10.1016/0168-9274(93)90036-q

Bond, T. G., & Fox, C. M. (2015). Applying the Rasch Model: Fundamental measurement in the human sciences (3rd ed.). London: Routledge.

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences* http://dx.doi.org/10.1007/978-94-007-6857-4_5

Brandes, J. A., McWhirter, P. T., Haring, K. A., Crowson, M. H., & Millsap, C. A. (2012). Development of the Indicators of Successful Inclusion Scale (ISIS): addressing ecological concerns. *Teacher Development,* 16(4), 463-488. http://dx.doi.org/10.1080/13664530.2012.717212

Byrne, B. M. (2010). *Structural equation modeling with AMOS: Basic concepts, applications and programming* (2nd ed.). New York: Routledge.

Cassidy TG, Jones RL and Potrac P (2008) *Understanding Sports Coaching: the Social, Cultural and Pedagogical Foundations of Coaching Practice*. New York: Routledge.

Costello, S., & Boyle, C. (2013). Pre-service Secondary Teachers' Attitudes Towards Inclusive Education. *Australian Journal of Teacher Education*, 38(4). http://dx.doi.org/10.14221/ajte.2013v38n4.8

DeVellis, R. (2003). *Scale Development: Theory and Application*. California: Sage publications.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research, 16*(1), 5-18. http://dx.doi.org/10.1007/s11136-007-9198-0

Ennis, C. (1992). Curriculum theory as practice: case studies of operationalized value orientations. *Journal of Teaching in Physical Education, 11*, 358-375. http://dx.doi.org/10.1123/jtpe.11.4.358

Ennis, C. (1994). Urban secondary teachers' value orientations: delineating curricular goals for social responsibility. *Journal of Teaching in Physical Education, 13*, 163-179. http://dx.doi.org/10.1123/jtpe.13.2.163

Ennis, C., & Chen, A. (1995). Teachers' value orientations in urban and rural school settings. *Research Quarterly for Exercise and Sport, 66*(1), 41-50. http://dx.doi.org/10.1080/02701367.1995.10607654

Ernst, C., & Rogers, M. R. (2009). Development of the Inclusion Attitude Scale for High School Teachers. *Journal of Applied School Psychology*, 25(3), 305-322. http://dx.doi.org/10.1080/15377900802487235

Evans, J. (2004). Making a difference? Education and 'ability' in physical education. *European Physical Education Review, 10*(1), 95-108. http://dx.doi.org/10.1177/1356336X04042158

Evans, J., & Davies, B. (1986). Sociology, schooling and physical education. In J. Evans (Ed.), *Physical education, sport and schooling: Studies in the sociology of physical education* (pp. 11-37). London: The Falmer Press.

Evans, J., & Davies, B. (1993). Equality, Equity and Physical Education. In J. Evans (Ed.), *Equality, Education and Physical Education* (pp.11-27). London: The Falmer Press.

Fitzpatrick, K. (2011). Trapped in the physical: Maori and Pasifica achievement in HPE. *Asia Pacific Journal of Health, Sport and Physical Education, 2*(3), 35-52. http://dx.doi.org/10.1080/18377122.2011.9730358

Flintoff, A. (2008). Targeting Mr Average: participation, gender equity and school sport partnerships. *Sport, Education and Society, 13*(4), 393-411. http://dx.doi.org/10.1080/13573320802445017

Flintoff, A., Fitzgerald, H., & Scraton, S. (2008). The challenges of intersectionality: Researching difference in physical education. *International Studies in Sociology of Education, 18*(2), 73-85. http://dx.doi.org/10.1080/09620210802351300

Gillespie, L. (2011). Exploring the 'how' and 'why' of Value Orientations in Physical Education Teacher Education. *Australian Journal of Teacher Education, 36*(9), http://dx.doi.org/10.14221/ajte.2011v36n9.4

Good, C., Aronson, J., & Harder, J. A. (2008). Problems in the pipeline: Stereotype threat and women's achievement in high-level math courses. *Journal of Applied Developmental Psychology, 29*(1), 17-28. http://dx.doi.org/10.1016/j.appdev.2007.10.004

Green, K., Smith, A., Thurston, M. and Lamb, K. (2007). Gender and secondary school National Curriculum Physical Education: change alongside continuity. In I. Wellard (Ed), *Rethinking gender and youth sport* (pp. 68-83). London: Routledge.

Hay, P. J., & lisahunter. (2006). "Please Mr Hay, what are my poss(abilities)?": legitimation of abililty through physical education practices. *Sport, Education and Society, 11*(3), 293-310. http://dx.doi.org/10.1080/13573320600813481

Hay, P., & Macdonald, D. (2010). Evidence for the social construction of ability in physical education. *Sport, Education and Society, 15*(1), 1-18. http://dx.doi.org/10.1080/13573320903217075

Hay, P. J., & Penney, D. (2013). *Assessment in physical education. A socio-cultural perspective.* Abingdon, Oxon.: Routledge

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1-55. http://dx.doi.org/10.1080/10705519909540118

Jeong, M., & Block, M. E. (2011). Physical Education Teachers' Beliefs and Intentions Toward Teaching Students With Disabilities. *Research Quarterly for Exercise and Sport*, 82(2), 239-246. doi: 10.1080/02701367.2011.1

Joreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & S. J. Long (Eds.), *Testing structural equation models* (pp. 294-316). Newbury Park: SAGE Publications.

Linacre, J. M. (2002a). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M. (2002b). What do infit and outfit, mean-square and standardize mean? *Rasch Measurement Transactions, 16*(2), 878.

Linacre, J. M. (2004). Optimizing Rating Scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch Measurement: Theory, models and application* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2015). *Facets computer program for many-facet Rasch measurement, version 3.71.4*. Beaverton, Oregon: Winsteps.com.

Lucas, P. (2011). Stereotyping *Ethics and Self-Knowledge* (Vol. 26, pp. 53-63): Springer Netherlands. http://dx.doi.org/10.1007/978-94-007-1560-8

Penney, D., & Evans, J. (2013). Who is physical education for? . In S. Capel & M. Whitehead (Eds.), *Debates in physical education* (pp. 157-170). Abingdon, Oxon.: Routledge.

Penney, D., & lisahunter (2006). Guest Editorial Overview: (Dis)Abling the (health and) physical in education: ability, curriculum and pedagogy. *Sport Education and Society, 11*(3), 205-209. http://dx.doi.org/10.1080/13573320600813358

Scanlan, J. N., Lannin, N. A., & Hoffmann, T. (2015). Can Rasch analysis enhance the abstract ranking process in scientific conferences? Issues of interrater variability and abstract rating burden. *Journal of Continuing Education in the Health Professions, 35*(1), 18-26. http://dx.doi.org/10.1002/chp.21263

Sharma, U., Loreman, T., & Forlin, C. (2012). Measuring teacher efficacy to implement inclusive practices. *Journal of Research in Special Educational Needs*, 12(1), 12-21. http://dx.doi.org/10.1111/j.1471-3802.2011.01200.x

Sirna, K., Tinning, R., & Rossi, T. (2008). The social tasks of learning to become a physical education teacher: considering the HPE subject department as a community of practice. *Sport, Education and Society, 13*(3), 285-300. http://dx.doi.org/10.1080/13573320802200636

Steele, C., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811. http://dx.doi.org/10.1037/0022-3514.69.5.797

Stone, J. (2002). Battling Doubt by Avoiding Practice: The Effects of Stereotype Threat on Self-Handicapping in White Athletes. *Personality and Social Psychology Bulletin, 28*(12), 1667-1678. http://dx.doi.org/10.1177/014616702237648

Tennant, A., & Pallant, J. F. (2006). Unidimensionality matters! (A tale of two Smiths?). *Rasch Measurement Transactions, 20*(1), 1048-1051.

Whitinui, P. (2005). The indigenous factor: The role of kapa haka asa culturally responsive learning intervention, *Waikato Journal of Education*, 10, 85-98.

Wilczenski, F. L. (1995). Development of a Scale to Measure Attitudes toward Inclusive Education. *Educational and Psychological Measurement*, 55(2), 291-299. http://dx.doi.org/10.1177/0013164495055002013

Wright, B. D., & Stone, M. H. (1999). *Measurement essentials* (2 ed.). Wilmington, DW: Wide Range, Inc.

Wright, J, and Burrows, L. (2006) Re-conceiving ability in physical education: a social analysis, *Sport Education and Society*, 11(3), 275-291. http://dx.doi.org/10.1080/13573320600813440