Journal of Research and Advances in Mathematics Education

Volume 6, Issue 3, July 2021, pp. 260 - 276
DOI: 10.23917/jramathedu.v6i3.14138
p-ISSN: 2503-3697, e-ISSN: 2541-2590

Introducing a measure of perceived self-efficacy for proof (PSEP): Evidence of validity

Benjamin Shongwe*, Vimolan Mudaly

Department of Mathematics Education, University of KwaZulu-Natal, South Africa

*Corresponding author: shongweb@ukzn.ac.za

ARTICLE INFO

Article history:
Received: 11 April 2021
Revised: 3 June 2021
Accepted: 4 June 2021
Published online: 27 July 2021
Published regularly: July 2021

Keywords:
Calibration, factor analyses, instrument development, self-efficacy for proof

ABSTRACT

It is widely recognized that students encounter difficulties with proof across all grades and beyond, yet standardized instruments related specifically to students' perceived self-efficacy for mathematical proof have not been readily available. The purpose of this study was to develop and investigate preliminary validity evidence for a new instrument for measuring self-efficacy for mathematical proof that can be of importance to the field. The new Perceived Self-Efficacy for Proof (PSEP) questionnaire is a self-administered, 8-item questionnaire that quantifies experimentation, conjecturing, inductive reasoning, justification, and validation. To validate the PSEP, two studies with 260 eleventh grade students—recruited from three Dinaledi schools in EThekwini metropolitan area, South Africa—were conducted. In Study 1 (n=128), face and content validity were evaluated, and an exploratory factor analysis (EFA) was performed. In Study 2 (n=132), a confirmatory factor analysis (CFA) was conducted and external validity was investigated. In both samples, the PSEP was found to possess good internal consistency reliability with relatively high factor loadings on a single component. Although the findings in this report represent preliminary validation evidence, it can be concluded that the PSEP is a valid, reliable and sensitive measure of 11th grade students' perceptions of their ability to construct a proof and may serve as a meaningful outcome in mathematical proof research and classroom proof education.

Introduction

It is widely recognized that students with higher levels of *self-efficacy*—also known as perceived ability, which denotes here the belief in one's abilities to succeed in a task—are more effortful, attempt more cognitively challenging problems, and persist longer in the face of difficulties in tasks (Oriol, Amutio, Mendoza, Da Costa, & Miranda, 2016). In addition, self-efficacy is a significant predictor of students' academic achievement (Daher, Gierdien, & Anabousy, 2021). This relationship is not new. For instance, Pajares and Kranzler (1995) investigated the influence of self-efficacy and mental ability of a sample of high school students engaging in mathematical problem solving tasks. They found that self-efficacy influenced both students' observed performance in the problem solving tasks and their mental ability with equal strength. However, traditional research on students' performances in proof has concentrated primarily on cognitive factors and less on affect

To cite this article:
Shongwe, B., & Mudaly, V. (2021). Introducing a measure of perceived self-efficacy for proof (PSEP): Evidence of validity. *JRAMathEdu (Journal of Research and Advances in Mathematics Education), 6*(3), 260-276. https://doi.org/10.23917/jramathedu.v6i3.14138

(Lau, Fang, Cheng, & Kwong, 2019). Yet, "[p]urely cognitive" behavior is extremely rare, and what is often taken for pure cognition is actually shaped – if not distorted – by a variety of factors (Schoenfeld, 1983, p. 330). Thus, an understanding of this construct (i.e., perceived self-efficacy for proof) could be of utmost importance to the field, particularly in drawing attention to the sources of students' difficulty with mathematical proof (hereinafter proof), a difficulty widely recognized as universal. However, existing validity studies of instruments designed to measure students' perceived ability to construct a proof remain unexplored despite the domain of Euclidean geometry demanding significantly different factors.

Broadly speaking, perceived self-efficacy for proof refers to a student's mechanism based on expectation to accomplish proof-related tasks. The purpose of the present study was threefold: distinguishing perceived self-efficacy from related constructs; developing an instrument designed to measure it; and investigating tentative validity evidence for it. The Perceived Self-Efficacy for Proof (PSEP) questionnaire is a brand new, self-administered, 8-item questionnaire operationalized by items describing the activities in which students engage in a meaningful proving process which characterizes the practices of research mathematicians: experimentation, conjecturing, seeking counterexamples, justification, and validation. To explore evidence of validity for the new measure of self-efficacy for proof, two studies were conducted.

Extensive research has been devoted to understanding the important place that the concept of self-efficacy occupies in the learning and instruction area. The most used and frequently cited instrument designed to measure high school students' self-efficacy is Betz and Hackett's (1993) Mathematics Self-Efficacy Scale. In addition, the researchers who have investigated students' performance in mathematics report similar effects of self-efficacy and the cognitive component.

In the past decade, large-scale international studies attempted to measures students' self-efficacy toward mathematics. For instance, Wilkins (2004) analyzed the Trends in International Mathematics and Science Study (TIMSS) data of several countries to understand the relationship between students' self-belief concepts in mathematics and respective achievement. He made contradictory findings; there were positive relationships at the student level, yet negative ones at the country level. A similar study undertaken by Shen and Tam (2008) found that whereas students in Korea and Singapore showed the highest achievement in TIMSS, they had lower self-efficacy than those in the Unites States (US).

Conceptual framing: self-efficacy

Conceptually, self-efficacy for proof is situated within Bandura's (1986) social cognitive perspectives of self-beliefs. The conceptual domain of self-beliefs constructs includes confidence, self-efficacy, self-concept, and anxiety (Morony, Kleitman, & Lee, 2012). These four constructs are substantively distinct. However, there is no consensus concerning their definitions and specificity in part due to the overlap between them, various perspectives from which these constructs have been investigated, and their operationalization and measurement (Bong & Skaalvik, 2003; Ferla, Valcke, & Cai, 2009; Hughes, Galbraith, & White, 2011). Whereas confidence statements are described as utterances intended to affect the belief of others about one's ability to provide correct answers to a task (Charness, Rustichini, & Van de Ven, 2018), self-efficacy examination "includes both an affirmation of a capability level and the strength of that belief" (Bandura, 1997, p. 282). The lack of consensus notwithstanding, it is possible to delineate an explicit distinction between confidence and self-efficacy.

According to Stankov, Lee, Luo, and Hogan (2012), confidence is an individual's state of being certain about successful performance in a task. Efklides (2011) sees confidence as a task-specific metacognitive experience in which an individual provides particular responses to indicate the level of certainty about the correctness of responses to a mathematical problem. For instance, an item such as "Mathematics is harder for me than for many of my classmates" in the TIMSS Students' Self-Confidence in Learning Mathematics index suggest a comparison between the student and his or her peers. Thus, measures of confidence are confined to making judgment about the correctness of an answer to a problem.

Now, let us turn to the distinguishing elements of the concept of self-efficacy. Central to the construct of self-efficacy is the distinction between perceived capability and observed accomplishment, respectively a non-cognitive component (judgment of capability) and an evaluative component (judgment of accomplishment). One other distinctive feature of self-efficacy is that no comparison of individuals with others takes place (Pajares, 2006). From these perspectives, it is reasonable to conclude that whereas confidence is viewed as a product, self-efficacy is taken as a process (learning and performing).

Calibration of self-efficacy

Calibration is the degree to which an individual student's judgment of his or her capability reflects the actual (observed) accomplishment in an objective task. Researchers have also reported that high achieving students have high self-efficacy and are relatively accurately calibrated in that they have more accurate self-perceptions (Zimmerman, 2009). In contrast, the notion of miscalibration (either over- or under-confident) arises when an individual's self-efficacy does not mirror the observed accomplishment (Alexander, 2013; Pajares, 2006). According to Chiu and Klassen (2010), this phenomenon (i.e., miscalibration) is associated with students in countries that are poor, culturally hierarchical, less more tolerant of uncertainty, or less flexible regarding gender roles; attributes broadly describing Felbrich, Kaiser, and Schmotz (2012) notion of collectivist cultures. Triandis' (1989) assertion that African cultures are collectivist was particularly important for this study. This assertion is important in light of Lawson' (2015) claim that in collectivist cultures, individuals need to adhere to group expectations which tend to be less tolerant of deviation from the norm.

An appreciation of miscalibration requires a reflective thought by the student. Baron, Gürçay, and Metz (2017) use the term "reflective thought" to denote moment in which one chooses "to be careful at the expense of speed" (p. 108). Thus, reflective thought relates to an individual's moment in which he or she ascribes meaning to the task at hand. Further, in the reflective thinking process, the basis of students' knowledge and beliefs are reviewed and evaluated (Kim & Silver, 2016). Thus, reflection is a process that can be seen as a means through which students learn mathematics meaningfully, including construction of proofs in ways typical of the practices of theoretical mathematicians. As Stylianides and Stylianides (2018) point out, these ways include seeking patterns, conjecturing, seeking counterexamples, proving, and justifying.

Significance of researching students' self-efficacy for proof

As already mentioned, despite the persistent and pervasive nature of students' difficulty with proof, limited attention is given to developing measures of self-efficacy for proof in the current literature. An assessment of students' self-efficacy for proof warrants attention as it could be of utmost importance for a variety of reasons. Understanding students' self-efficacy for proof could draw attention to the sources of students' difficulty with the concept of proof. Students' self-efficacy for proof may be a beneficial construct to employ when investigating correlations of students' difficulty with the concept of proof. Task-specificity assessments can provide prediction indexes and insights not available from general assessments of influence of self-efficacy on observed performance (Seegers & Boekaerts, 1996).

Conceptualizing and measuring students' self-efficacy for proof may also be useful for understanding their overall achievement in mathematics. Given the central role which self-efficacy plays in the exercise of personal agency by its strong impact on thought, affect, motivation, and action, calibration may also be one way students metacognitively observe their academic progress (Bandura, 2014). Metacognitive monitoring entails the evaluation of self-efficacy relative to task demands to assess whether the standards being pursued are attainable or beyond a student's reach. However, as Oriol et al. (2016) point out, in human endeavors, failures do not necessarily lower aspirations. Final but not least, miscalibration could be unraveled. The next section provides a brief background on the theoretical model of self-efficacy for proof.

Research methods

Current research

The present study consisted of two phases (i.e., Study 1 for scale development and Study 2 for testing the reliability of the scale with an independent sample) to give credence to the reliability of the scale (Field, 2017). In Study 1, the definition of students' self-efficacy for proof was used to generate an initial set of items. These items were based on Dreyfus and Hadas' (1987) principles of proof understanding whose outline is outside the scope of the present study. It was hypothesized that the items on the scale would indeed load onto a single factor and thus enable an examination of self-efficacy for proof as a unidimensional factor.

In Study 2, additional evidence of validity was the factor structure obtained by conducting a confirmatory factor analysis (CFA) and examining external validity. It was hypothesized that the single-factor solution would provide good fit to the data and that the students' PSEP scale would be empirically distinct from related constructs. In addition to providing the methodology of this study, the next section documents the development of the measurement instrument and its psychometric properties in line with the principle that prior to being utilized, an instrument must be rigorously assessed for its scores to be labeled valid and reliable (reproducible).

Study 1: Participants and procedure

The PSEP measurement instrument was administered to 128 grade 11 students conveniently recruited from three Dinaledi schools in EThekwini metropolitan area, South Africa. Dinaledi schools followed a mandatory mathematics curriculum that includes algebra, trigonometry, analytical geometry, and Euclidean geometry in addition to financial mathematics, statistics and probability. The poor performance of previously disadvantaged students in mathematics and physical science has been of national concern and a priority for South African government, as evidenced in the establishment of the Dinaledi School Project in 2001 (Department of Basic Education [DBE], 2009).

Of the 128 students participating in the study, 56 were male, 67 were females, and 5 did not report their gender. The ages of the participants ranged from 15 to 22 (M = 16.26, SD = 1.44). These data were collected during February of 2017. In addition, the participants selected multiple ethnic/racial categories in which 76 identified themselves as Black, 41 as

Asian, 4 as Colored, and 4 as "Other." Three participants chose not to report ethnicity. Exploratory factor analysis (EFA) was performed using data obtained from this sample.

Study 2: Participants and procedures

Participants were 132 students who were in the second half of the larger sample recruited to determine whether the items and their factor were the same across two independent samples (Brown, 2014). In this sample, there were 56 males, 71 were females, and 5 did not report their ages. The ages of the participants ranged from 14 to 22 (M = 16.96, SD = 1.04). The racial/ethnic composition of the sample was Black (52.3%), Asian (41.7%), Colored (3.8%), and "Other" (2.3%). Three participants did not indicate their race or ethnicity. The procedure for Study 2 mirrored that of Study 1.

Prior to describing item development process and factor analyses methods in detail, it is worth reminding the reader that the purpose of the present study was to (1) develop a measure of students' self-efficacy for proof and (2) determine the psychometric properties of this instrument. The measure was developed after a thorough review of the proof literature, self-efficacy instruments, expert judgments, and cognitive interviews with respondents. The next sections describe these studies, in turn. Specifically, the sections demonstrate a "hybrid" approach in which an exploratory factor analysis (EFA) is performed to help generate a new theory and following up on its results using a CFA with separate data to test the theory (Hair, Black, Babin, & Anderson, 2019).

Study 1: Item development and content and face validation

Based on the preliminary definition of students' self-efficacy for proof, an initial pool of 13 items was developed. Content validity was evaluated by soliciting feedback from five experts (two faculty colleagues, two doctoral students, and one postdoctoral researcher) with knowledge of the concepts of self-efficacy and mathematical proof. However, one doctoral student invited to comment on content validity of the instrument did not respond. One of the experts has specific expertise in scale development and another has published extensively on the subject matter (i.e., proof).

In the examination of content validity, some items which were deemed (*a priori*) as very similar were included so that the experts might help in determine a more appropriate way to phrase them. The experts also evaluated face validity, a component of content validity pertaining to whether the test "looks valid" to the participants (Creswell, 2018). Following Adelson and McCoach's (2011) and McCoach, Gable, and Madura's (2013) description of the best practices for the development of new, valid, and reliable scales, the experts rated three aspects of content validity: agreement if each item measured self-efficacy for proof or a different construct; certainty for categorizing the items as self-efficacy for proof (1 = not very sure, 2 = pretty sure, 3 = very sure); and the extent to which an item is relevant for measuring self-efficacy for proof (1 = *low/no relevance*, 2 = *mostly relevant*, 3 = *highly relevant*).

Next, cognitive interviews with five respondents whose characteristics were similar to the target population were conducted to refine and assess item interpretation. Specifically, further face validation took place through conducting two rounds of cognitive interviewing with these respondents. Questions were carefully considered for gender neutrality, clarity, brevity, interpretability, and subsequent translations into other languages (Creswell, 2018).

Results and Discussion

Results

An analysis of experts' feedback provided both qualitative and quantitative data. From a qualitative perspective, they commented on individual items and provided feedback in relation to the appropriateness of the operational definition, construct coverage, item phrasing (clear and unambiguous), and appropriate level of the scale for use with high schoolers. The definition was revised to acknowledge two critical aspects.

First, it was to acknowledge that students' self-efficacy for proof is indeed a matter of their own perceptions rather than an objective fact. Second, the definition was revised to acknowledge that students also lack an appreciation of the meaning of the concept of proof in mathematics. Specifically, the revision was prompted by respondents' inability to distinguish between the common use of the term "proof" and its technical meaning in the mathematics discipline. As a consequence, self-efficacy for proof was redefined as a construct that describes student's naïve or informed and perceived competence in successfully accomplishing a proof construction task; hence the questionnaire is hereinafter referred to a Perceived Self-Efficacy for Proof (PSEP). On the one hand, "naïve" means viewing proof as merely a means to verify the truth of a proposition using few cases. On the other hand, "informed" denotes viewing proof as serving not only the function verification but also explanation, communication, discovery, and systematization (de Villiers, 1990).

Although one item and the response categories were somewhat modified by the experts, the theoretical concepts essentially remained the same. Only two rounds are reported here because relatively few new insights emerged from the third round (Willis, 2005). Based on the feedback responses from cognitive interviews, one item was dropped for lack of clarity. In addition, modification to grammar and the word choice were also made.

From a quantitative perspective, the percentage of experts who agreed that the items operationalized the construct of perceived self-efficacy for proof ranged from 85.7% to 100.0%. Items with over 80% agreement (9 items) were retained for further consideration. Mean certainty and mean relevance ranged from 2.50 to 3.00 and from 2.14 to 3.00, respectively. The agreement rate for the item ("Seek counterexamples") that had under 85.7% was, however, retained because it was considered a key feature in proof construction. One more item ("Verify the truth of a proposition") was deleted due to its low relevance rating (2.14) and multiple expert comments about its redundancy ($M_{\text{certainty}}$ = 2.49, $M_{\text{relevance}}$ = 2.55). Based on the feedback from the expert at scale development, midpoint anchor label "Uncertain" was removed to better assess miscalibration; it was replaced with "Neither yes or no" so as to maintain a seven-point Likert scale. The final version of PSEP contained 8 items along with 3 demographic questions (i.e., age, ethnicity or race, and gender). The advantage of using a multiple-item PSEP scale lies in its ability to measure internal consistency reliability and thus reflect the "cohesiveness" of different items within a unique factor (Creswell, 2018).

Exploratory factor analysis 1

The questionnaire was presented on the same paper as the proof construction task in which participants were expected to experiment with few triangles (plane), conjecture, seek counterexamples, construct, and validate a proof (however, on different pages). The participants were asked to self-report on the PSEP questionnaire whose stem was, "Do you think you have the ability to do the following task?" They responded on a seven-point Likert response scale: 1 = *strong no*, 2 = *no*, 3 = *somewhat no*, 4 = *neither yes or no*, 5 =

somewhat yes, 6 = *yes*, and 7 = *strong yes*. At the start of the administration of the survey, participants reported on demographic data. Then, directions on the instrument informed participants that they were to express their perceived ability to engage in the list of different activities related to proof construction. On average, the completion of the survey took 15 minutes after which they were directed to turn to the next page which contained the proof construction task.

As depicted in Figure 1, participants were provided with a few triangles with which observe and seek a pattern, generalize, conjecture, seek counterexamples, prove, and validate the correctness of the proof. The figure also shows possible activities (observation, generalization, and conjecturing). The proving task took on average 25 minutes to complete. Participants completed the survey and the task after normal school hours.

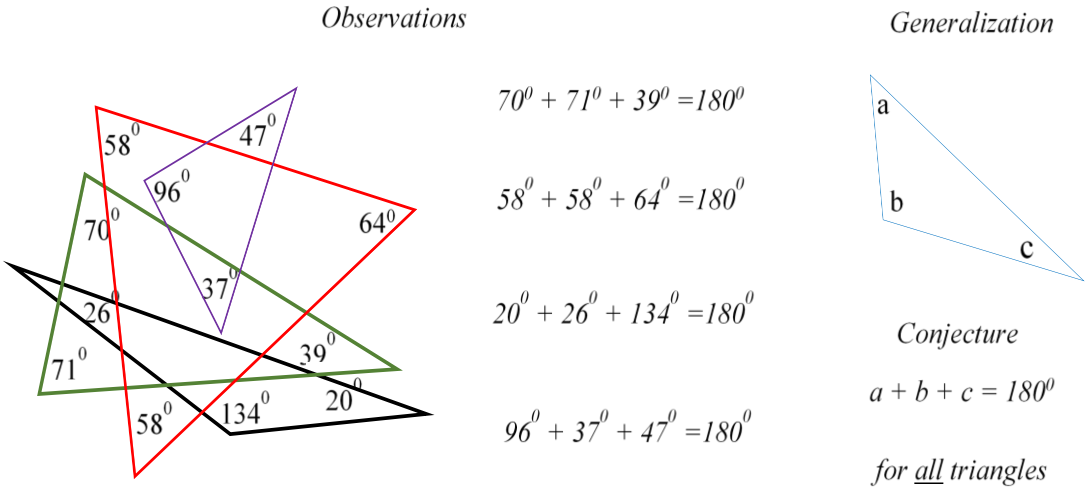

Figure 1. Observation of few cases resulting in a generalization about angles on a plane triangle

Exploratory factor analysis 2

Two separate SPSS extraction methods used were principal axis factoring (PAF) and principal components analysis (PCA). In an attempt to identify the factors that underlie the structure of the PSEP, PAF as an extraction method was initially performed because it is the "classic factor analytic approach" (Pett, Lackey, & Sullivan, 2003, p. 103). Item 5 ("Give up in the face of difficulties") in Table 1 was reverse coded so that all items were in the same direction and consisted of positive correlations. In addition to the fact that the determinant score was above the rule which indicated an absence of multicollinearity, the Barlett's test of sphericity (Chi-square = 413.42, *df* = 28, *p*< .001) which rejected the null hypothesis that the correlation matrix is an identity matrix, and the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy which was closer to 1 (.838) suggested the data were suitable for EFA (Field, 2017).

In summarizing the variables into fewer components, parallel analysis (PA), which Hayton, Allen, and Scarpello (2004) describe as "one of, if not the, most accurate method for determining the number of factors to retain" (p. 197) was considered together with the Kaiser-Guttman rule of retaining eigenvalues greater than 1.0, and the "elbow joint" of the scree plot. The observed eigenvalues exceeded the comparison (i.e., "parallel") eigenvalues as derived from randomly permutated observed data having the same sample size and items as the original dataset thus supporting the single-factor structure observed in scree plot. Thus, the eight items loaded onto one factor to create a construct than can explain the interrelationships among these items.

Factor analysis and internal consistency reliability results

The results of EFA identified the factor structure and other items for fine tuning. A PAF extraction provided pattern and structure coefficients as well as factor correlations. Items with pattern coefficients of at least .40 were retained (McCoach et al., 2013; Pett et al., 2003). The pattern matrix of the PAF solution with oblique rotation, structure matrix, and communalities are presented in Table 1.

Given that the factor explained 65.90% of the common variance, it is reasonable to suggest that the PSEP scale consisted of items that operationally defined the construct of perceived self-efficacy for proof. The resulting internal consistency of the items within the PSEP scale (i.e., Cronbach's coefficient alpha) in this sample, which represents the proportion of variability attributable to the true score (Pett et al., 2003), was strong (α = .92). Cronbach's coefficient alpha is defined as the degree to which the set of items in the scale co-vary relative to their sum score (Raykov & Marcoulides, 2011). Although an alpha coefficient of .70 is regarded as an acceptable threshold for reliability, .80 and .95 are preferred for the psychometric quality of scales (Hair et al., 2019). The first two columns in Table 1 present oblimin-rotated pattern and structure matrices following PAF extraction of the single factor. In the last column, means and standard deviations for these items are presented. Factor loadings (i.e., the strength of the relationship that each item has with the factor) were significantly high.

Table 1
The PSEPscale items and their coefficients

Item		Coefficients			
		Pattern	Structure	(h^2)	*M* (*SD*)
1.	Engage in experimentation to seek patterns	.48	.49	.44	4.70 (0.94)
2.	Make a conjecture	.88	.79	.63	4.65 (0.73)
3.	Verify if the conjecture using few cases	.63	.67	.45	4.73 (1.06)
4.	Seek counterexamples	.63	.73	.54	4.45 (1.00)
5.	Give up in the face of difficulties*	.74	.74	.58	4.37 (0.99)
6.	Use previously proven statements	.73	.58	.59	4.00 (1.05)
7.	Formally write out each step of the proof	.86	.65	.79	4.22 (0.94)
8.	Examine the proof for accuracy	.52	.47	.46	3.63 (1.09)

Note: Pattern coefficient values less than .20 were suppressed; h^2 = communalities of the measured items using PCA. Oblimin rotation was used. *Reverse-coded item scores after the item reverse coding

Interpretation of PSEP factor

The PSEP measures students' perceived self-efficacy for proof. High scale scores reflect a perception that a student holds in relation to accomplishment in a proof construction task and low scores reflect a perception that a student holds in relation to accomplishment in a proof construction task. Having reverse-coded item 5 of the PSEP scale, item scores were averaged to obtain a scale score ranging from 0 to 7. Overall, scores reflected participants' varying degrees in which they perceived themselves to have the ability to accomplish a proof construction task (*M* = 3.90, *SD* = .76).

The total scores in the PSEP scale ranged from 0 to 48; the sample contained participants whose perception that they could hardly construct a proof was strong and those who felt strongly that they could successfully construct a proof. Also, the proportion of variance in the items accounted for by the factor was strong for all the items loading on it. Further, a fairly good variability for responses on each individual item in PSEP was observed. As Hair et al. (2019) propose, given that the dataset had several high factor loading scores (> .80), then the sample size was reasonably sufficient.

The total variance explained by the self-efficacy for proof construct was at the threshold (49.99%) which suggested a reasonably good factor solution. In social science research, unlike in the natural sciences where information is often more precise, extracted factors usually explain only 50% to 60% (Hair et al., 2019). Put another way, the single extracted component (factor) explained nearly 50% of the variability in the eight items on SEP. Having developed this theory using EFA, the next section uses CFA to hypothesize an *a priori* model of the underlying factor structure and examines if this model fits the data adequately (Bandalos, 1996).

Confirmatory factor analysis: Dimensionality investigation

A CFA was performed on the second half of the sample to test dimensionality (i.e., to assess the fit of the data to the theoretical structure). This approach was adopted to maintain independence of the EFA and CFA samples. However, prior to estimating how well the model fit the hypothesized structure, the *P-P* plot was examined to determine the shape of the distribution of data; that is, whether the data approximated a normal distribution. *P-P* plots compare the cumulative probability of observed data with a theoretically ideal normal distribution (Field, 2017). As displayed in Figure 2, in addition to absence of outliers, the data values in the *P-P* plot lay along the diagonal line. This suggested that the data approximately resembled a normal distribution; thus, performing and interpreting the linear regression model was appropriate for the data.

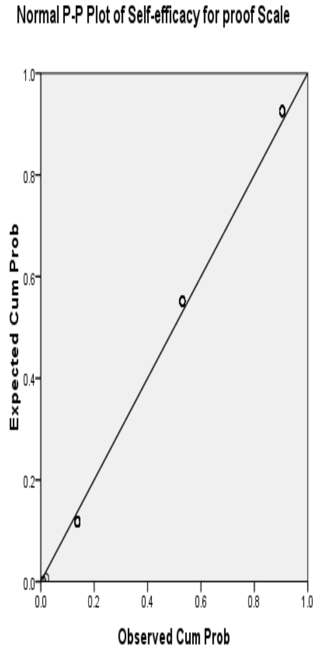

Figure2. Standardized normal P-P Plot for the PSEP data

The CFA was performed with the use of SPSS analysis of moment structures (AMOS) because there were no missing data. The CFA model consisted of a single latent variable, representing the singe factor derived from the results of EFA (i.e., perceived self-efficacy for proof). The test for the hypothesized unidimensional structure obtained from a satisfactory initial factor analysis in Study 1 followed systematic fit assessment procedures in line with recommendations by Creswell (2018). The fit statistics that were selected and evaluated in this study were Chi-square test of exact fit, the comparative fit index (CFI), the Tucker-Lewis Index (TLI), and the root mean square error of approximation (RMSEA).

CFA model and internal consistency reliability

The nonsignificant Chi-square value suggested a good fitting model. In assessing the extent to which the hypothesized model fit the actual data, Hair et al's. (2019) standards were adopted: evidence of reasonable fit included obtaining a nonsignificant χ^2 value, a normed χ^2 (i.e., χ^2/df) value below 5, comparative fit index (CFI) and incremental fit index (IFI) values of .90 or higher, and root mean square error of approximation (RMSEA) value of .08 or less. Although these statistics indicate how well the model fits the data, the strength of the structural paths in the model is determined by the squared multiple correlations (SMC) (Hair et al., 2019).

Squared multiple correlations in Table 2 suggested the absence of singularity (i.e., SMC close to 0.00) and multicollinearity (i.e., SMC close to 1.00). In addition, standardized regression weights were determined to compare directly the relative contribution of each item on the factor (Hair et al., 2019). The one-factor solution with the eight observed PSEP items comprised standardized regression weights (factor loadings) which were statistically significant at p< .01.

As shown in Table 2, all the standardized regression weights and the squared multiple correlations were high. For instance, standardized regression weight of item 1 ("Engage in experimentation to seek patterns") on the factor it represents (self-efficacy) was higher than the threshold of .70. Also, the squared multiple correlation (.71) of this item suggested that 71% of its variance is accounted for by self-efficacy for proof construct while the remaining 29% of its variance is accounted for by measurement error. In short, self-efficacy for proof explained half (50%) of the variation in the item with the other half being error variance. Taken together, self-efficacy for proof is theoretically meaningful and contributes appreciably to account for the variance in the dataset. Put another way, these results suggested that self-efficacy for proof is indeed a construct that defines a distinct cluster of interrelated items. As a consequence, all the eight items in EFA were retained.

The Chi-square for the model suggested a good model fit:χ^2 (17) = 27.108, p > .05. The other alternative fit indices also indicated a good fit to the data: normed χ^2 = 1,60, CFI = .98, TLI = .96, and, RMSEA = .07. Put another way, the results suggested a reliable model for future forecasts. Similar to Study 1, the internal reliability consistency was high (α = .91). Similar to Study 1, participants in Study 2 varied in terms of the extent to which they were certain of their proof construction ability (*M* = 4.32, *SD* = 1.24). In addition to descriptive statistics for each item, Table 3 shows that there was relatively also good variability observed for each item in the PSEP scale.

Table 2

Standardised regression weights (factor loadings) and squared multiple correlations (communalities) for the one-factor CFA model (*n* = 132)

Item	Factor loadings	Squared multiple correlations	
1. Engage in experimentation to seek patterns	.74	.20	4.52 (1.37)
2. Make a conjecture	.75	.56	4.05 (1.43)
3. Verify if the conjecture is true using few cases	.70	.48	4.32 (0.96)
4. Seek counterexamples	.71	.51	4.30 (1.65)
5. Persevere in the face of difficulties*	.74	.55	4.44 (0.69)
6. Use previously proven statements	.67	.38	4.15 (1.35)
7. Formally write out each step of the proof	.68	.46	4.78 (1.29)
8. Examine the proof for accuracy	.79	.26	4.03 (1.17)

External validity

Further evidence of validity was provided by assessing construct validity through examining the correlations between PSEP and other measures. Confirmatory factor analysis has gained the most prominence as an alternative model to simultaneously evaluate both discriminant and convergent validity to probe construct validity (Boateng, Neilands, Frongillo, Melgar-Quiñonez, & Young, 2018) of the conceptualization of students' self-efficacy. The purpose of investigating convergent validity is to estimate the relationship between the self-efficacy for proof construct and a similar construct in which stronger correlation coefficients would suggest support for convergent validity (Creswell, 2018). In contrast, the purpose of discriminant validity is to estimate the relationship between scale scores and distinct constructs in which weaker correlation coefficients would suggest support for discriminant validity (Creswell, 2018). Specifically, convergent validity was examined by correlating PSEP scores with the four-item mathematics rigor scale. Discriminant validity was assessed by correlating PSEP scores with the measure of perceived confidence. The next section describes the measures used to assess external validity of the PSEP scale.

Mathematics rigor

Although TIMSS is hypothesized as low-stakes tests—that is, tests without consequences for the student but potentially high-stakes for schools and countries which may be judged deficient if scores are poor (Jackson, Khavenson, & Chirkina, 2020)—its measure of mathematics rigor provided an excellent opportunity to test students' perceptions. A student who rarely, if ever, thought about the rigor of mathematical problems would have a low score on the instrument. The participants responded to 4 items on the scale (e.g., "I need to invest greater effort in mathematics" and "I learn things quickly in mathematics (or science)" based on a 5-point Likert scale (1 = *strongly disagree*, 2 = *disagree*, 3 = *neither agree nor disagree*, 4 = *agree*, 5 = *strongly agree*). The item, "There is no need to study harder in mathematics", was reverse-coded so that all items were in the same direction. The total score for each participant ranged from 4 to 16, and this score was then divided by 4 to obtain an average score for rigor. TIMSS reported a Cronbach's α coefficient of .94. A similarly excellent internal consistency reliability α coefficient of .81 was obtained in this study.

An assessment of this construct in participants was deemed important following Shen and Tam's (2008) report that students in low-performing countries, for example, South Africa and Morocco, are likely to say that they perceive mathematics (including proof) as being easy and that they learn the subject matter quickly. An examination of correlations coefficients of PSEP scores and those of the four-item measure of perceived mathematics rigor provided additional evidence of convergent validity. Put another way, strong coefficients between perceived self-efficacy and rigor were evidence in support of convergent validity.

Confidence

Given that Stankov et al. (2012) found that confidence alone explained the major part of variance on the self-beliefs constructs including self-efficacy, it was used as a measure to determine discriminant validity. Students' confidence is typically assessed by asking participants to indicate, on a percentage scale, how confident they are that their just-provided responses to the proving task are correct (Morony et al., 2012). According to Bandura (1997), asking participants to express their confidence in solving a mathematical problem (in this case, constructing a proof) similar to that presented in a cognitive test or examination serves to increase prediction of academic outcomes. He points out that the expression of confidence can provide prediction indices and insights not available from broader assessments of self-efficacy.

Participants rated the strength of their confidence to successfully accomplish the activity on the PSEP scale using a five-point Likert scale ranging from 1 (*not confident at all*) to 5 (*completely confident*). The total score for each participant ranged from 8 to 32. This score was then divided by 8 to obtain an average self-efficacy for proof score. An excellent internal consistency reliability α coefficient of .92 was obtained.

Table 3 models the interrelationships among items by representing them with three (latent) variables. An examination of correlations coefficients of PSEP scores and those of the confidence measure provided additional evidence of discriminant validity (i.e., perceived self-efficacy for proof is empirically distinct from confidence). The relationship between perceived self-efficacy for proof and confidence was weak, an indicator of discriminant validity. Put differently, the results indicated that although these two constructs were related, they were conceptually distinct. In contrast, the high and statistically significant correlation coefficient between PSEP and mathematics rigor demonstrated that self-efficacy for proof and rigor were related constructs, an indicator or convergent validity. Correlations between PSEP and each of the three external measures are provided in Table 3.

Table 3
Correlations among measures for Study 2

Measure	1	2	3
SEP	–		
Rigor	.79**	–	
Confidence	.20**	.28**	–

Note: Correlation is significant at the 0.01 level (2-tailed).

The correlation matrix of all the scales shows interrelationships in a pattern that supports their construct validity. On the one hand, ratings were significantly correlated within each item across the participants, evidencing convergent validity. On the other hand, ratings were not significantly correlated with each other, providing evidence for discriminant validity. Thus, it can be claimed that this PSEP measurement scale has construct validity as demonstrated by the results of both convergent and discriminant validity. The results were consistent with the hypothesis that there exist weak correlations between PSEP and the measure of confidence. The results were also consistent with the hypothesis that there exists a strong relationship between the PSEP and the measure of rigor. Considered together, these findings provided tentative evidence for convergent and divergent validity. Further, these results suggested that PSEP is a unidimensional measure comprising multiple items that capture the underlying domain of students' perceived self-efficacy for proof.

Discussion

Findings from the present study represent an important first step in conceptualizing and measuring high school students' perceived self-efficacy for proof. In Study 1, a rigorous content-evaluation process was performed to establish evidence of content validity for the PSEP scale. The exploratory factor analysis of the scale yielded a single-factor structure on which all items loaded highly. Also, an excellent internal consistency reliability index was obtained. Similar to Study 1, a new independent sample in Study 2 yielded scores that demonstrated high internal consistency reliability and good model fit (i.e., the items on the scale can be used to predict or explain students' self-efficacy for proof). Further, this sample demonstrated convergent validity with the measure of rigor and discriminant validity with the measure of confidence. The CFA results show that students' self-efficacy for proof can be understood in terms of eight variables.

Scores on PSEP for both Study 1 and Study 2 samples demonstrated good variability. Self-efficacy for proof is a valuable metacognitive monitoring process in a learning environment because it helps students to correct distorted self-efficacy. Self-efficacy for proof also provides useful diagnostic information for teachers and schools by comparing the confidence score to accuracy (i.e., the percentage of correct answers in a task) to examine confidence judgments (Moore & Healy, 2008).

Additional findings were made concerning miscalibration. The finding that participants tended to overestimate their self-efficacy for proof relative to actual achievement was similar to Chiua and Klassen's (2010) findings. They found that miscalibration was a feature of students in countries that were poor, less egalitarian, or less flexible regarding gender roles. Vancouver and Kendall (2006) have shown that overestimation of one's self-efficacy is characteristic of students in a collectivist society, which Hofstede (1986; 2011) defines as the tendency within a culture toward gregariousness and group orientation (e.g., South Africa; Triandis, 1989). Overestimation may result in less preparation, less help-seeking, and ultimately poor achievement (Kim & Silver, 2016). According to Triandis (1989), urban samples tend to be individualistic, and traditional-rural samples tend to gravitate toward collectivism within the same culture. The samples in this study were urban thus suggesting that urban societies tend to embrace individualism rather than collectivism. Thus, the results related to miscalibration are inconsistent with Triandis' (1989) findings. One explanation for this contradictory result is that, although African cultures are considered collectivist, the sample consisted of participants from other ethnic groups whose culture could be different.

Limitation

The results reported in this study must be treated with caution. The process of validating the PSEP scale was complicated by the absence of appropriate known-to-be reliable measures of self-efficacy for proof. The use of other measures is regarded as the best option available. Although the design of this study incorporated an attempt to reduce the threats to external validity of the results by using two studies with different participants from the population of eleventh graders because random selection could not be done, some caution on inference of generalizability must be made. Specifically, some students may already have been harboring intentions to change to the other mathematics whose content excludes proofs (e.g., mathematical literacy) due to career choices, parental disapproval, peer pressure, and so on. Because of these factors, it is difficult to generalize the results of this study to other contexts.

The fact that only Dinaledi high schools were sampled is considered a limitation. Students in these schools may have demonstrated evidence of self-efficacy for proof that may be unique. Therefore, caution must be exercised in generalizing the findings of this study to the general high school population because conclusions regarding a global view of self-efficacy in proof-related tasks were not made. However, given the sparse but needed research on perceived self-efficacy for proof, it is critically important to understand competence beliefs in schools with a focus on mathematics and physical sciences in light of the budget spending allocated to these schools. Also, direct input from the participants in the questionnaire should not be underestimated; such input can broaden our understanding of the difficulties that students encounter when learning proof. It is through this understanding that will guide us to new methods that make the learning and teaching of proof meaningful. In addition, based on the operative definition of self-efficacy for proof, the statements in the scale represent substantively valid measures.

Conclusion

In conclusion, previous studies have focused on self-efficacy for mathematics in general. The present study has considered self-efficacy specifically for proof. The purpose of this study was to develop and psychometrically validate an instrument specifically designed to measure students' perceived self-efficacy for proof in high schoolers and thus add to the literature on understanding students' difficulty with proof which exists across national and cultural boundaries. The PSEP instrument was developed and validated with two samples utilizing EFA and CFA.

Measuring students' self-efficacy for proof is important because proof construction is considered a vital practice for understanding mathematics and how its knowledge is constructed. For instance, measuring students' self-efficacy for proof may help in the identification and assistance of those with high self-self-efficacy scores but underachieve (i.e., miscalibration). This study offers insight into the challenges and skills current school leaders must face to leverage technology to improve student learning. Embarking on an instrument development design was intended to provide the mathematics education field with a tool to make sound generalizations about calibration.

On the basis of tentative validity evidence in the present study, PSEP scale has demonstrated to be a metacognitive tool to measure and diagnose students' distorted self-efficacy for proof perceptions. Thus, measuring students' perceived self-efficacy for proof is particularly important for the improvement and deepening of our understanding of the mathematical domain of proof which persistently continues to be an area of great difficulty for many students. The findings of this study would be further strengthened through replication with reliable measures of self-efficacy for proof. Future research efforts may be directed at calibration (using bias scores) to compare subgroups such as gender or age, examination of demonstrative geometry resources (e.g., dynamic geometry software), and ethnic differences. Policy decisions can be guided by research efforts focusing on such comparisons.

Acknowledgment

The authors deeply appreciate the participation of learners in this study. The study was supported in part by the University Capacity Development Programme (UCDP) grant at the University of KwaZulu-Natal (UKZN). Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the UCDP.

Bibliography

Adelson, J. L., & McCoach, D. B. (2011). Development and psychometric properties of the Math and Me Survey: Measuring third through sixth graders' attitudes toward mathematics. *Measurement and Evaluation in Counseling and Development, 44*(4), 225-247. http://doi.org/10.1177/0748175611418522

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*, 1-3. http://doi.org/10.1016/j.learninstruc.2012.10.003

Bandalos, B. (1996). Confirmatory factor analysis. In J. Stevens (Ed.), *Applied multivariate statistics for the social sciences* (3rd ed., pp. 389-420). Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Bandura, A. (2014). Social cognitive theory of moral thought action. In W. M. Kurtines, & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development* (pp. 69-128). Hillsdale, NJ: Psychology Press.

Baron, J., Gürçay, B., & Metz, S. E. (2017). Reflective thought and actively open-minded thinking. In M. E. Toplak, & J. Weller (Eds.), *Individual differences in judgment and decision making from a developmental context* (pp. 107-126). New York, NY: Routledge.

Betz, N. E., & Hackett, G. (1993). *Mathematics self-efficacy scale.* (Mental Measurements Yearbook 14, No. 14081939). Abstract retrieved from Mental Measurements Yearbook.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health, 6:149*. http://doi.org/10.3389/fpubh.2018.00149

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review, 15*, 1-40.

Brown, T. (2014). *Confirmatory factor analysis for applied research.* New York, NY: Guilford Press.

Charness, G., Rustichini, A., & Van de Ven, J. (2018). Self-confidence and strategic behavior. *Experimental Economics, 21*(1), 72-98.

Chiua, M. M., & Klassen, R. M. (2010). Relations of mathematics self-concept and its calibration with mathematics achievement: Cultural differences among fifteen-year-olds in 34 countries. *Learning and Instruction, 20*, 2-17.

Creswell, J. W. (2018). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (6th ed.). Boston, MA: Pearson.

Daher, W., Gierdien, F., & Anabousy, A. (2021). Self-efficacy in creativity and curiosity as predicting creative emotions. *Journal of Research and Advances in Mathematics Education, 6*(2), 86-99. http://doi.org/10.23917/jramathedu.v6i2.12667

de Villiers, M. D. (1990). The role and function of proof in mathematics. *Pythagoras, 24*, 17-24.

Department of Basic Education [DBE]. (2009). *The Dinaledi Schools Project: Report from a strategic engagement between the national department of education and business on increasing support for mathematics and science in education in schools.* Pretoria: Department of Basic Education.

Dreyfus, T., & Hadas, N. (1987). Euclid may stay—and even be taught. In M. M. Lindquist, & A. P. Shulte (Eds.), *Learning and teaching geometry, K-12* (pp. 47-58). Reston, VA: National Council of Teachers of Mathematics.

Efklides, A. (2011). Interactions of metacognition with motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist, 46*, 6-25.

Felbrich, A., Kaiser, G., & Schmotz, C. (2012). The cultural dimension of beliefs: An investigation of future primary teachers' epistemological beliefs concerning the nature of mathematics in 15 countries. *ZDM Mathematics Education, 44*, 355-366.

Ferla, J., Valcke, M., & Cai, Y. (2009). Academic self-efficacy and academic self-concept: Reconsidering structural relationships. *Learning and Individual Differences, 19*, 499-505.

Field, A. (2017). *Discovering statistics using IBM SPSS statistics* (5th ed.). London: Sage.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). New Jersey, NJ: Pearson.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*, 191-205.

Hofstede, G. (1986). Cultural differences in teaching and learning. *International Journal of Intercultural Relations, 10*, 301-320.

Hofstede, G. (2011). Dimensionalizing cultures: The Hofstede model in context. *Online Readings in Psychology and Culture, 2*(1), 26 pages.

Hughes, A., Galbraith, D., & White, D. (2011). Perceived competence: A common core for self-efficacy and self-concept? *Journal of Personality Assessment, 93*, 278-289. http://doi.org/10.1080/00223891.2011.559390

Jackson, M., Khavenson, T., & Chirkina, T. (2020). Raising the stakes: Inequality and testing in the Rtussian education system. *Social Forces, 98*(4), 1613-1635. http://doi.org/10.1093/sf/soz113

Kim, Y., & Silver, R. E. (2016). Provoking reflective thinking in post observation conversations. *Journal of Teacher Education, 67*(3), 203-219. http://doi.org/10.1177/0022487116637120

Lau, Y., Fang, L., Cheng, L. J., & Kwong, H. K. (2019). Volunteer motivation, social problem solving, self-efficacy, and mental health: A structural equation model approach. *Educational Psychology, 39*(1), 112-132.

Lawson, D. M. (2015). *Family violence: Explanations and evidence-based clinical practice.* New York: Wiley.

McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain. School and corporate applications* (3rd ed.). New York, NY: Springer.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*(2), 502-517.

Morony, S., Kleitman, S., & Lee, Y. P. (2012). Predicting achievement: Confidence vs self-efficacy, anxiety, and self-concept in Confucian and European countries. *International Journal of Educational Research, 58*, 79-96.

Oriol, X., Amutio, A., Mendoza, M., Da Costa, S., & Miranda, R. (2016). Emotional creativity as predictor of intrinsic motivation and academic engagement in university students: The mediating role of positive emotions. *Frontiers in Psychology, 7:1243*. http://doi.org/10.3389/fpsyg.2016.01243

Pajares, M. F. (2006). Self-efficacy during childhood and adolescence. In M. F. Pajares, & T. Urban (Eds.), *Adolescence and education: Self-efficacy and adolescence* (Vol. 5, pp. 339-367). Greenwich, Connecticut: Information and Age.

Pajares, M. F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology, 20*, 426-443.

Pett, M. A., Lackey, N. R., & Sullivan, J. J. (2003). *Making sense of factor analysis: The use of factor analysis for instrument development in health care research.* Thousand Oaks, CA: Sage.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory.* New York, NY: Routledge.

Schoenfeld, A. H. (1983). Beyond the purely cognitive: Beliefs systems, social cognitions, and metacognitions as driving forces in intellectual performance. *Cognitive Science, 7*, 329-363.

Seegers, G., & Boekaerts, M. (1996). Gender-related differences in self-referenced cognitions in relation to mathematics. *Journal for Research in Mathematics Education, 27*(2), 215-240.

Shen, C., & Tam, H.-P. (2008). The paradoxical relationship between student achievement and self-perception: A crossnational analysis based on three waves of TIMSS data. *Educational Research and Evaluation, 14*(1), 87-100.

Stankov, L., Lee, J., Luo, W., & Hogan, D. J. (2012). Confidence: A better predictor of academic achievement than self-efficacy, self-concept and anxiety? *Learning and Individual Differences, 22*, 747-758.

Stylianides, A. J., & Stylianides, G. J. (2018). Addressing key and persistent problems of students' learning: The case of proof. In A. J. Stylianides, & G. Harel (Eds.), *Advances in mathematics education research on proof and proving: An international perspective* (pp. 99-113). Cham, Switzerland: Springer.

Triandis, H. C. (1989). The self and social behaviour in differing cultural contexts. *Psychological Review, 96*, 506-520.

Vancouver, J. B., & Kendall, L. N. (2006). When self-efficacy negatively relates to motivation and performance in a learning context. *Journal of applied psychology, 91*(5), 1146-1153. http://doi.org/10.1037/0021-9010.91.5.1146

Wilkins, J. L. (2004). Mathematics and science self-concept: An international investigation. *The Journal of Experimental Education, 72*(4), 331-346.

Willis, G. B. (2005). *Cognitive interviewing: A tool for improving questionnaire design.* Thousand Oaks, CA: Sage.

Zimmerman, B. J. (2009). Self-efficacy and educational development. In A. Bandura (Ed.), *Self-efficacy in changing societies* (pp. 202-231). Cambridge, UK: Cambridge University Press.