



BRIEF REPORT

An Instrument for Measuring Critical Appraisal Self-Efficacy in Rheumatology Trainees

Juliet Aizer,¹  Erika L. Abramson,² Jessica R. Berman,¹ Stephen A. Paget,¹ Marianna B. Frey,³ Victoria Cooley,² Ying Li,² Katherine L. Hoffman,² Julie A. Schell,⁴ Michael D. Tiongson,⁵ Myriam A. Lin,⁶ and Lisa A. Mandl¹ 

Objective. Self-efficacy, the internal belief that one can perform a specific task successfully, influences behavior. To promote critical appraisal of medical literature, rheumatology training programs should foster both competence and self-efficacy for critical appraisal. This study aimed to investigate whether select items from the Clinical Research Appraisal Inventory (CRAI), an instrument measuring clinical research self-efficacy, could be used to measure critical appraisal self-efficacy (CASE).

Methods. One hundred twenty-five trainees from 33 rheumatology programs were sent a questionnaire that included two sections of the CRAI. Six CRAI items relevant to CASE were identified a priori; responses generated a CASE score (total score range 0–10; higher = greater confidence in one's ability to perform a specific task successfully). CASE scores' internal structure and relation to domain-concordant variables were analyzed.

Results. Questionnaires were completed by 112 of 125 (89.6%) trainees. CASE scores ranged from 0.5 to 8.2. The six CRAI items contributing to the CASE score demonstrated high internal consistency (Cronbach's $\alpha = 0.95$) and unidimensionality. Criterion validity was supported by the findings that participants with higher CASE scores rated their epidemiology and biostatistics understanding higher than that of peers ($P < 0.0001$) and were more likely to report referring to studies to answer clinical questions (odds ratio 2.47, 95% confidence interval 1.41–4.33; $P = 0.002$). The correlation of CASE scores with percentage of questions answered correctly was only moderate, supporting discriminant validity.

Conclusion. The six-item CASE instrument demonstrated content validity, internal consistency, discriminative capability, and criterion validity, including correlation with self-reported behavior, supporting its potential as a useful measure of critical appraisal self-efficacy.

INTRODUCTION

Critical appraisal is an important skill in evidence-based practice. This process, which involves “assessing and interpreting evidence by systematically considering its validity, results and relevance,” is particularly important in rheumatology, which has seen a rapid emergence of both new diagnostic tools and therapeutic approaches (1). Physicians who can competently assess

and interpret the results of clinical research will be best equipped to use this information to appropriately inform patient care.

To critically appraise the medical literature, physicians require not only the knowledge and skills to perform critical appraisal but also sufficient motivation to do so. Social cognitive and self-determination theories provide useful lenses through which to consider and analyze motivation (2–4). These theories identify self-efficacy as an important factor influencing behavior.

The study was supported by the Weill Cornell Medicine Clinical and Translational Science Center NIH grant UL1-TR-000457. Dr. Aizer's work was supported by the Nanette Laitman Education Scholar Award in Entrepreneurship. Dr. Aizer, Ms. Frey, Dr. Tiongson, Ms. Lin, and Dr. Mandl's work was supported by the Hospital for Special Surgery Academy of Medical Educators.

¹Juliet Aizer, MD, MPH, Jessica R. Berman, MD, Stephen A. Paget, MD, FACP, FACR, Lisa A Mandl, MD, MPH: Weill Cornell Medicine and Hospital for Special Surgery, New York, New York; ²Erika L. Abramson, MD, MSc, Victoria Cooley, MS, Ying Li, MS, Katherine L. Hoffman, MS: Weill Cornell Medicine, New York, New York; ³Marianna B. Frey, BA: University of Rochester School of Medicine and Dentistry, Rochester, New York; ⁴Julie A. Schell, EdD, MS, BS:

The University of Texas at Austin, and Harvard University, Cambridge, Massachusetts; ⁵Michael D. Tiongson, MD: University of Rochester Medical Center, Rochester, New York; ⁶Myriam A. Lin, BA: Hospital for Special Surgery, New York, New York.

Author disclosures are available at <https://onlinelibrary.wiley.com/action/downloadSupplement?doi=10.1002%2Facr.2.11505&file=acr211505-sup-0001-Disclosureform.pdf>.

Address correspondence via email to Juliet Aizer, MD, MPH, at Aizerj@hss.edu.

Submitted for publication May 27, 2022; accepted in revised form September 12, 2022.

Self-efficacy theories fall into two categories. The first is general self-efficacy, which refers to self-perceptions of one's ability to succeed across diverse domains or situations. The second is task-specific self-efficacy, or self-perceptions of one's ability to perform a specific task successfully (3).

Although there is debate as to whether general self-efficacy is a qualitatively different construct than self-esteem, self-efficacy (both general and specific) as a concept correlates with a range of achievement outcomes, including academic and job performance. Supporting this, research has shown that people are more likely to engage in activity-specific behavior when they believe they will be successful at executing that activity (5). To optimize learning, it is important to be able to identify whether a learner has relative deficiencies in both actual competence and specific self-efficacy (6). Therefore, to promote the use of critical appraisal in clinical practice, graduate medical education programs must develop not only knowledge and skills but also specific self-efficacy related to performing critical appraisal. The lack of a tool to measure critical appraisal self-efficacy (CASE), however, limits assessment of the impact of educational programs on this construct.

Specific self-efficacy scales evaluate individuals' perception regarding the degree to which they believe they can perform activities relevant to the task at hand (5). Although some have suggested a role for generalized or global self-efficacy as a mediator of specific self-efficacy, more specific measures of self-efficacy have been found to have greater precision relative to domain-specific tasks in addition to performing better than generalized instruments for explaining and predicting performance in a particular context (3). Because we are interested in the specific domain of critical appraisal, we focus on specific self-efficacy in this study.

Because the development of valid outcome measures can be extremely time consuming and resource intensive, when feasible, it is often more efficient to identify instruments with evidence of validity from prior work that can be applied or modified rather than creating new instruments *de novo* (7).

The Clinical Research Appraisal Inventory (CRAI) is a 92-item instrument measuring self-efficacy related to activities involved in conducting clinical research (8). Each item on the CRAI has a response option from 0 to 10 (no confidence to total confidence). The CRAI was rigorously developed, including assessments and refinements to ensure the language was clear and understandable. It has accrued evidence for its validity when used to measure clinical research self-efficacy in pre- and postdoctoral trainees, trainees in Master of Science programs, and academic physicians (8–10). Many sections of the CRAI relate to activities involved in performing clinical research (for example, "Funding a Study" and "Protecting Research Subjects and Responsible Conduct of Research") that, although extremely important to investigators, are less relevant to most rheumatology trainees who are not pursuing research careers. However, two sections ("Designing a

Study" and "Interpreting Data") include questions focused on critical appraisal. We sought to assess the feasibility and validity of using a subset of questions from the CRAI to measure CASE in rheumatology trainees.

MATERIALS AND METHODS

Selection of CASE items. Two investigators involved in teaching critical appraisal to trainees in rheumatology (LAM, JA) and a third investigator with teaching critical appraisal experience in undergraduate and graduate medical education (ELA) each independently reviewed the CRAI to identify items relevant to CASE. These investigators were instructed to identify questions that evaluated any aspect of critical appraisal. The choice of questions was informed by a systematic review from the Cochrane Database that described the potential benefits of critical appraisal "in interpreting studies, informing them of potential biases, increasing comprehension of numerical results, and helping them to decide whether articles are relevant, valid and how they should influence the care of their patients" (1). Questions that had the potential to measure self-efficacy as it relates to each of these behaviors were chosen (Table 1).

Study population. In September 2017, all Accreditation Council for Graduate Medical Education (ACGME)-accredited rheumatology programs ($n = 135$) were invited to enroll their trainees in a web-based curriculum to build epidemiology and biostatistics knowledge and skills (Hospital for Special Surgery Critical Literature Assessment Skill Support in Rheumatology [HSS CLASS-Rheum[®]]) (11). We developed HSS CLASS-Rheum as a question-based tool to support rheumatology trainees in learning knowledge and skills relevant to critical appraisal (11). This convenience sample of enrolled trainees received a preprogram questionnaire anonymously eliciting responses to demographic questions, including self-reported gender, race, and ethnicity, and two sections from the CRAI ("Designing a Study" and "Interpreting Data," which included the six questions relevant to CASE). The two sections of the CRAI were administered in a standardized format that maintained the original item order and instructions. In June 2018, at the end of the academic year during which they had access to HSS CLASS-Rheum, participants received a similar post-program questionnaire. We used data from these questionnaires to evaluate the performance of questions from the CRAI in these rheumatology trainees.

Statistical analysis. Descriptive statistics were derived using percentages, means, standard deviations (SDs), and minimums and maximums as appropriate. Internal consistency of the six items selected from the CRAI was assessed using Cronbach's α . To assess content validity, their dimensionality was assessed through parallel analysis and exploratory factor

Table 1. Behaviors relevant to critical appraisal addressed by CASE items

	We would like to know how confident you are that you can successfully perform these tasks today					
	CASE Item 1: compare major types of studies (such as case reports; case-control, cross-sectional, longitudinal, and epidemiological studies; clinical trials, etc)	CASE Item 2: recognize important threats to internal and external validity applicable to each research design	CASE Item 3: state the purpose, strengths, and limitations of each study design	CASE Item 4: explain the outcome of a given analysis in terms of the originally stated hypotheses or research questions	CASE Item 5: express appropriate methodological and theoretical cautions in interpreting results	CASE Item 6: identify limitations of a study
Interpreting studies	✓			✓	✓	
Recognizing potential biases		✓	✓			✓
Comprehending numerical results				✓		
Determining study relevance	✓	✓				✓
Determining study validity		✓	✓		✓	✓
Determining how articles should influence patient care	✓	✓	✓			✓

Abbreviation: CASE, critical appraisal self-efficacy.

analysis. Criterion validity, which includes retrospective, concurrent, and predictive validity, was assessed by analyzing the correlation of CASE scores with participants' 1) self-reported frequency of referring to an original study when faced with a clinical question (retrospective validity), 2) description of their understanding of epidemiology and biostatistics relative to other fellows (concurrent validity), and 3) number of HSS CLASS-Rheum questions attempted (predictive validity). Construct validity was assessed by analyzing the correlation of postprogram CASE scores with the percentage of HSS CLASS-Rheum questions answered correctly out of the total number of questions answered. Pearson's or Spearman's rank correlation coefficient was used to describe the relationship between CASE scores and other measures, as appropriate. The chi-square test was used to compare the difference in the proportions of reporting prior coursework in epidemiology or biostatistics between groups of participants. Two-sample independent *t*-tests or analysis of variance with the Holm adjustment for multiple comparisons was used to compare mean CASE scores between groups of participants as appropriate. A paired *t*-test was used for comparison of the mean pre- and postprogram CASE scores for participants who completed both pre- and postprogram CASE questions. Bivariate

logistic regression was used to determine the association of baseline CASE scores with self-reported behavior. All *P* values were two-sided, and statistical significance was evaluated at the 0.05 level. Statistical analyses were performed in R Version 3.5.1 (12). The open-source R packages used in the analysis include summarytools, psy, psych, and stats (Supplementary Table 2). Research was in accordance with the Helsinki Declaration. Exemption was obtained by the Hospital for Special Surgery Institutional Review Board.

RESULTS

Identifying CASE questions from the CRAI. Three investigators (LAM, ELA, and JA) independently identified the same six items from the CRAI as having direct relevance to CASE (Table 1). A seventh item was identified as possibly relevant by one investigator ("Choose an appropriate research design that will answer a set of research questions and/or test a set of hypotheses"), but after discussion, consensus was reached that this item was framed in terms of performing research rather than critical appraisal and was not included. Questions retained included major content areas of CASE discussed in the Cochrane systematic review

(Table 1). A CASE (composite critical appraisal self-efficacy) score was defined as the unweighted average of an individual's responses to these six questions.

Study population characteristics. Thirty-three rheumatology programs from across the United States enrolled 125 trainees in HSS CLASS-Rheum (Supplementary Table 1). One hundred eighteen trainees submitted preprogram questionnaires; 112 (95%) answered all CASE items. Thirty-five percent were male, and the majority were White and non-Hispanic. The group had similar gender, race, and ethnicity, as self-reported by US rheumatology trainees overall (13) (Table 2).

Table 2. Characteristics of participants answering CASE questions (N = 112)

Characteristic	Number (%)
Female	73 (65.2%)
Male	39 (34.8%)
African American or Black	4 (3.6%)
American Indian or Alaska Native	0 (0%)
Asian or Indian subcontinent	34 (30.4%)
Native Hawaiian or other Pacific Islander	0 (0%)
White, Caucasian, or Middle Eastern	62 (55.4%)
Other	12 (12.5%)
Hispanic or Latinx	10 (8.9%)
Fellowship	
Year 1	53 (47.3%)
Year 2	47 (42.0%)
Year 3	10 (8.9%)
Year 4	2 (1.8%)
Prior pediatrics training	28 (25.0%)
Prior coursework in epidemiology or biostatistics	63 (56.3%)
Self-reported understanding of epidemiology and biostatistics compared with other rheumatology trainees	
Far greater understanding	0 (0%)
Somewhat greater understanding	6 (5.4%)
Average understanding	61 (54.5%)
Somewhat less understanding	34 (30.4%)
Far less understanding	11 (9.8%)
Self-reported frequency of referral to original studies when faced with a clinical question	
Never	1 (0.9%)
Rarely	8 (7.1%)
Sometimes	51 (45.5%)
Often	44 (39.3%)
Always	8 (7.1%)
Factors identified as reducing referral to studies	
Do not have time	64 (57.1%)
Do not always have access to original articles	31 (27.7%)
Do not know how to interpret original articles	25 (22.3%)
Do not know how to search for original articles	18 (16.1%)
Do not need original articles to treat my patients well	4 (3.6%)
Other	8 (7.1%)

Abbreviation: CASE, critical appraisal self-efficacy.

Description of baseline data. There was a range of responses to each CASE item (Figure 1). Composite CASE scores ranged from 0.5 to 8.2, with a mean of 4.18 (SD = 1.76). Cronbach's α for the six CASE items was 0.95, indicating strong internal consistency. An exploratory factor analysis to assess content validity found the six CASE items to be unidimensional and captured a total item variance of 74.8% (root mean square error of approximation = 0.242, root mean square residual = 0.05, Tucker-Lewis Index = 0.844; Supplementary Figure 1).

Participants with higher preprogram CASE scores were more likely to report referring to studies to answer clinical questions (odds ratio 2.47, 95% confidence interval 1.41-4.33, $P = 0.002$). Mean preprogram CASE scores were also higher in participants who rated their understanding of epidemiology and biostatistics as greater than that of their peers. Those who rated their understanding as somewhat greater, average, somewhat less, or far less than that of their peers had mean CASE scores of 6.53, 4.61, 3.55, and 2.48, respectively ($P < 0.0001$). This stepwise, dose-dependent relationship supports concurrent validity. Preprogram CASE scores also predicted the number of HSS CLASS-Rheum questions participants attempted ($P < 0.001$), supporting predictive validity. However, preprogram CASE scores did not predict the percentage of attempted HSS CLASS-Rheum questions answered correctly, despite the fact that higher CASE scores were seen in those with previous epidemiology or biostatistics coursework (mean 4.76 vs. 3.44, $P < 0.001$).

Comparisons between pre- and postprogram data.

Forty-three participants (38%) submitted the optional postprogram questionnaire; 41 of 43 (95%) answered all six CASE items on both pre- and postprogram questionnaires. Those who did complete the post-program questionnaire were of similar gender, race, ethnicity, year of rheumatology training, and prior pediatric training to those who did not complete it. However, those who completed the post-program questionnaire reported more prior coursework in epidemiology or biostatistics (73% vs. 46%, $P = 0.01$) and had higher mean CASE scores at baseline than those who did not (4.8 vs. 3.8, $P = 0.003$). In the 41 participants who completed the CASE questions pre- and post program, mean CASE scores increased from 4.8 preprogram to 6.3 post program ($P < 0.0001$), demonstrating responsiveness of the CASE instrument. Although the increase in mean CASE scores was greater for participants with preprogram mean CASE scores in the lowest versus highest quartile (mean increase 2.25 vs. 0.63, $P = 0.01$), the postprogram mean CASE scores remained significantly different (postprogram mean of 4.90 for lowest quartile and 7.52 for highest quartile, $P = 0.001$). There was only a moderate correlation between the percentage of

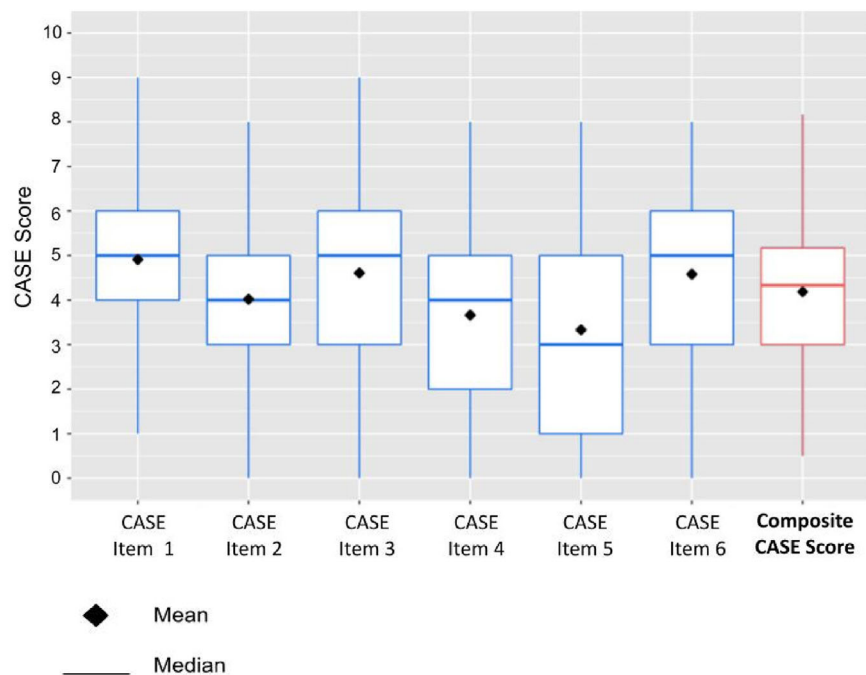


Figure 1. Responses to individual critical appraisal self-efficacy (CASE) items and composite CASE scores (N = 112).

attempted questions answered correctly in the HSS CLASS-Rheum learning modules and postprogram CASE scores (Spearman's $r = 0.39$, $P = 0.011$). This suggests the CASE questions do not simply measure knowledge of biostatistics and epidemiology, supporting discriminant validity.

DISCUSSION

To our knowledge, there is no instrument with evidence for its validity to measure CASE. Given that specific self-efficacy is an important factor influencing behavior, it would benefit rheumatology program directors to be able to effectively assess not only trainees' ability, but also their specific self-efficacy for appraising the rheumatic disease literature.

Three experts identified items from the CRAI with high relevance to CASE (1) (Table 1) and explored the performance of these six items when administered to rheumatology trainees. Cronbach's α of 0.95 indicates strong internal consistency, and the exploratory factor analysis indicates a unidimensional structure. The broad range of CASE scores suggests discriminative capability. These six items, focused specifically on CASE, are quick to administer. This is in contrast to the much longer 92-item CRAI, which includes additional domains relevant to the performance of clinical research (e.g., "Terminate a collaboration that isn't working" or "Locate appropriate forms for a grant application") that are not related to critical appraisal.

The moderate correlation between postprogram CASE scores and percentage correct on HSS CLASS-Rheum questions suggests CASE scores measure something different from demonstrated knowledge, providing evidence for discriminant validity.

This is consistent with the understanding that perceived self-efficacy is a belief that does not always correlate tightly with actual knowledge or capability. Self-efficacy can be overreported in those lacking mastery over a particular domain and underreported despite high performance on domain-specific tasks (8,14,15). Rather than simply reflecting content knowledge, self-efficacy is an attitude that can influence behavior. Indeed, the fact that prior to taking the course, participants with higher CASE scores were more likely to report they refer to studies to answer clinical questions suggests a link between CASE and a relevant behavior, providing evidence of criterion validity. After participation in HSS CLASS-Rheum, mean CASE scores increased, indicating responsiveness. The finding that participants with higher CASE scores attempted more questions in HSS CLASS-Rheum further supports criterion validity; however, this conclusion is tempered by the fact that participation in HSS CLASS-Rheum could have been affected by external factors, such as fellowship program requirements.

This study has some limitations. The postprogram response rate was 38%, potentially limiting the generalizability of our results. This moderate response rate was unsurprising given that responding to the postprogram questionnaire was an optional activity for trainees. Though responders were demographically representative of the entire cohort, they had higher baseline CASE scores, and a greater proportion reported prior coursework in epidemiology and biostatistics, which may have introduced bias in the exploratory pre-post program analysis. This study only evaluated rheumatology trainees in a clinical research-focused epidemiology and biostatistics course; these participants may not be representative of all rheumatology trainees or learners in other

disciplines. Future studies are needed to further explore discriminant validity, to assess how CASE scores perform in relation to generalized measures of self-efficacy regarding critical appraisal behaviors, and to help tease apart the various factors driving associations between CASE scores and critical appraisal behaviors.

This study also has many strengths. We selected items from a validated instrument and administered items in a format that maintained the original item order and instructions. We included a social science measurement expert as well as physicians with expertise in both epidemiology and educational psychology pedagogy in the item selection process to ensure content validity. We sampled trainees from a large number of diverse programs and had pre-post program data with which to explore the prospective performance of CASE scores.

In conclusion, we identified six items from the CRAI and evaluated their performance as a composite measure of CASE in rheumatology trainees. CASE scores demonstrated evidence of validity based on expert opinion, internal structure, and relation to other variables and behaviors. Whether these six CASE items can be administered as a stand-alone instrument to measure critical appraisal self-efficacy should be validated in other cohorts.

ACKNOWLEDGMENTS

The authors wish to thank the trainees and training programs in rheumatology who participated in HSS CLASS-Rheum, as well as the HSS Academy of Medical Educators; without their support this project would not have been possible.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be published. Dr. Aizer had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Aizer, Abramson, Berman, Paget, Frey, Cooley, Schell, Tiongson, Lin, Mandl.

Acquisition of data. Aizer, Frey, Cooley, Tiongson, Mandl.

Analysis and interpretation of data. Aizer, Abramson, Frey, Cooley, Li, Hoffman, Schell, Lin, Mandl.

REFERENCES

1. Horsley T, Hyde C, Santesso N, et al. Teaching critical appraisal skills in healthcare settings [review]. *Cochrane Database Syst Rev* 2011; CD001270.
2. Bandura A. The explanatory and predictive scope of self-efficacy theory. *J Soc Clin Psychol* 1986;4:359–73.
3. Bandura A. *Self-efficacy: the exercise of control*. New York: W. H Freeman; 1997.
4. Deci EL, Vallerand RJ, Pelletier LG, et al. Motivation and education: the self-determination perspective. *Educ Psychol (Lond)* 1991;26: 325–46.
5. Phillips JC, Russell RK. Research self-efficacy, the research training environment, and research productivity among graduate students in counseling psychology. *Couns Psychol* 1994;22:628–41.
6. Zimmerman BJ. Self-efficacy: an essential motive to learn. *Contemp Educ Psychol* 2000;25:82–91.
7. Artino AR Jr. Academic self-efficacy: from educational theory to instructional practice. *Perspect Med Educ* 2012;1:76–85.
8. Mullikin EA, Bakken LL, Betz NE. Assessing research self-efficacy in physician-scientists: the clinical research appraisal inventory. *J Career Assess* 2007;15:367–87.
9. Lipira L, Jeffe DB, Krauss M, et al. Evaluation of clinical research training programs using the clinical research appraisal inventory. *Clin Transl Sci* 2010;3:243–8.
10. Mills BA, Caetano R, Rhea AE. Factor structure of the clinical research appraisal inventory (CRAI). *Eval Health Prof* 2014;37:71–82.
11. Aizer J, Schell JA, Frey MB, et al. Learning to critically appraise rheumatic disease literature: educational opportunities during training and into practice. *Rheum Dis Clin N Am* 2020;46:85–102.
12. R Core Team. R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018. URL: <https://www.R-project.org/>.
13. Academy for Academic Leadership. 2015 workforce study of rheumatology specialists in the United States. URL: <https://www.rheumatology.org/portals/0/files/ACR-Workforce-Study-2015.pdf>.
14. Kruger J, Dunning D. Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *J Pers Soc Psychol* 1999;77:1121–34.
15. Pajares F. Gender differences in mathematics self-efficacy beliefs. In: Gallagher AM, Kaufman JC, editors. *Gender differences in mathematics: an integrative psychological approach*. Cambridge (UK): Cambridge University Press; 2005. p. 294–315.