BMC Psychology

RESEARCH

Open Access

Uncovering potential interviewer-related biases in self-efficacy assessment: a study among chronic disease patients

Magdalena Holter[1*], Alexander Avian[1], Martin Weger[2], Sanja Strini[2], Monja Michelitsch[2], Victoria Winkler[1], Agnes M. Kloft[1,3], Julia Groß[1], Thomas Falb[2], Maximilian Gabriel[2], Manuel Großpötzl[2], Andreas Wedrich[2] and Andrea Berghold[1]

Abstract

Background Self-efficacy refers to an individual's belief in their ability to accomplish specific tasks and achieve goals, and plays an essential role in achieving positive outcomes in a wide range of domains. Central to the measurement of any form of self-efficacy is the assessment without bias, also in case of an interview situation.

Methods Outpatients with macular edema, an eye disease, participated in this questionnaire-based cross-sectional study. The study assessed self-efficacy using the General Self-Efficacy Scale (GSE) in German. Interviewers read questionnaires aloud to patients. Differential item functioning (DIF) was investigated using likelihood-ratio $\chi 2$ tests for interviewer, sex, age, education, working status, income, diagnosis, and health-status.

Results The analysis included $N = 556$ patients. Median age was 68.4 (IQR: 62.0 – 76.0) years and mean overall GSE score 32.8 (SD: 4.81). No DIF was detected for interviewer. However, DIF was found in item 1 for education (uniform DIF, $NCDIF_{\text{no degree vs. degree}} = 0.042$; easier with degree vs. none), in item 1 and 3 for income (item 1: non-uniform DIF, $NCDIF_{<€\,1{,}125\ \text{vs.}\geq€\,1{,}125\leq€\,1{,}950} = 0.050$ / $NCDIF_{<€\,1{,}125\ \text{vs.}\geq€\,1{,}950} = 0.099$; item 3: uniform DIF, $NCDIF_{<€\,1{,}125\ \text{vs.}\geq€\,1{,}125—\leq€\,1{,}950} = 0.024$ / $NCDIF_{<€\,1{,}125\ \text{vs.}\geq€\,1{,}950}$: 0.095; both easier with higher income), in item 2 for working status (uniform DIF, $NCDIF_{\text{retired vs. other}} = 0.017$; easier if working) and in item 3 for sex (non-uniform DIF, $NCDIF_{\text{male vs. female}} = 0.043$; easier for women in low ability, harder for them from medium ability on).

Conclusions Given that no DIF was detected concerning interviewers, our findings indicate that an objective assessment of self-efficacy in a face-to-face interview may be feasible, provided that interviewers receive appropriate training. Since DIF effects concerning other patients characteristics found were small, the GSE may provide a relatively bias free way to assess self-efficacy in an interview setting.

Keywords General self-efficacy scale, Interviewer bias, Administration mode, Item response theory, Differential item functioning

Background

Self-efficacy refers to an individual's belief in their ability to accomplish specific tasks and achieve goals, and plays an essential role in achieving positive outcomes in a wide range of domains [1–3]. It is a crucial concept in various fields, including medicine, psychology, and education [4–7].

*Correspondence:
Magdalena Holter
magdalena.holter@medunigraz.at
[1] Institute for Medical Informatics, Statistics and Documentation, Medical University of Graz, Graz, Austria
[2] Department of Ophthalmology, Medical University of Graz, Graz, Austria
[3] Department of Computer Science, Aalto University, Espoo, Finland

Self-efficacy is typically measured using questionnaires that assess an individual's confidence in performing particular behaviors or tasks. Since self-efficacy can be very specific in various fields, there are distinct scales in different domains [8]. For example, in healthcare, specific measures have been developed for various conditions such as diabetes [9] musculoskeletal rehabilitation [10], childbirth [11], pain management [12], and preventive health services [13]. In contrast, there are also questionnaires designed more general that can be used in different populations and situations (e.g. General Self-Efficacy Scale (GSE) [14]).

Central of measuring any form of self-efficacy is the assessment without bias [15]. A robust operationalization of self-efficacy is needed [15]. Moreover, it is emphasized that self-efficacy items should be tailored to specific domains and not influenced by other constructs [16, 17]. To investigate a potential bias in self-efficacy measures, within the framework of item response theory (IRT), differential item functioning (DIF) can be used [18]. DIF occurs when different groups of people with the same underlying level of self-efficacy have different probabilities of giving a certain response to a specific item on the scale. This can lead to biased results when comparing groups, e.g. women always scoring lower on a particular item despite having the same level of self-efficacy as men. The presence of items with DIF is a severe threat to the validity of the measure for self-efficacy and to the conclusions based on the scores resulted from the items with DIF.

DIF of the GSE has been investigated in several studies and contradictory evidence has been found. While some investigations reported the presence of DIF concerning sex [19, 20], others did not observe such differences [21–23]. Similarly, certain studies identified DIF related to age among specific items [19, 24, 25], while others failed to detect such contrasts [21, 22]. Furthermore, DIF was noted to be associated with education in one study [24], yet it remained absent in others [21, 22]. Additionally, a study revealed DIF concerning work status [24]. These ambiguous findings underscore the importance of assessing and addressing DIF when using the GSE, ensuring measurement invariance across diverse populations or administration modes.

Essentially, there are two distinct modes for filling out questionnaires: those that include the interaction with an interviewer, such as face-to-face interviews, and those that can be completed without the need of an interviewer, such as self-administration [26]. The presence of an interviewer introduces a new potential source of DIF [27]. A disadvantage is the potential for interviewer bias, where the interviewer's influence, whether intentional or unintentional, may affect how individuals respond to the questionnaire [27].

For example, interviewers can differ in their ability to maintain a neutral appearance or intonation [28]. Intonations in words often gives additional meaning to speech [29]. Indirect communication and tone can affect cognitive and emotional processing of individuals, while body language provides essential cues for controlling one's own social appearance [26]. The characteristics and behaviors that respondents attribute to an interviewer can have an impact on the answers given [30]. A further issue in interviews can be an increase in positive or socially desirable responses. This can happen due to the interviewer's characteristics or because respondents might hesitate to disclose beliefs they think the interviewer might not share [31].

At present, there is limited research on interviewer-induced DIF [32–34]. Literature focused on the influence of other administration modes, such as paper–pencil, web-based or interactive voice system [35–38]. However, as the population ages and the importance of questionnaire accessibility grows, it becomes important to investigate whether specific questionnaires can be seamlessly integrated into interview settings without compromising their psychometric properties. The comparability of GSE scores across different interviewers and diverse groups when administered via face-to-face interviews remains an open question.

The aim of the study was to investigate the German GSE in patients with a chronic eye disease affecting vision and reading ability with respect to DIF in relation to interviewers. Additionally, sex, age, education, working status, income, type of macular edema, diagnoses and health status were investigated regarding DIF.

Methods

Study design

This questionnaire-based cross-sectional study consists of an ad-hoc sample from the population of outpatients from the Department of Ophthalmology of the Medical University of Graz. Data were collected from March 2020 until the end of February 2022. The ethical committee of the Medical University of Graz approved the study (32–101 ex 19/20). This is a secondary analysis of a recently published study [34].

Participants

Included patients suffered from the chronic disease macular edema, an accumulation of fluid in the macula, which impairs vision and can lead to severe visual impairment. Patients were included in this study if they had a macular edema due to diabetes or retinal vein occlusion, were at least 18 years old and spoke German well

enough, in order to understand the questionnaires. It was also necessary that the patients' hearing abilities were at a level where it was possible to communicate verbally with them. Patients were not included if they had a macular edema due to other causes or suffered from cognitive impairment.

Data collection

First, informed consent was signed by participating patients. The questionnaire was completed with the help of interviewers between medical eye examinations, as patients were unable to read, due to their eye condition and the application of dilating eye drops. All four interviewers were trained in the interview process and outcome measures, resulting in a standardized process to ensure the objectivity and comparability of the interviews. The training included familiarization with the study material, study procedure and a rehearsal interview, in which different potential situations with patients and the according reactions were simulated. This took about four hours. Additionally, first interviews at the outpatient clinic were overseen.

Several questionnaires were administered, the GSE was the sixth questionnaire. The web-based pseudonymization tool 'iPSN' [39] was used for the pseudonymization of participants. Answers from participants were gathered and stored in LimeSurvey [40].

Outcome measures

The GSE [41] was used to assess self-efficacy. The items were administered orally, in accordance with the German version of the questionnaire. It consists of 10 items which are answered on a four-point Likert-type response scale from "Not at all true" to "Exactly true". For example, one item states "I can solve most problems if I invest the necessary effort". Answers are summed up to a score. Reliability was found to be adequate with Cronbach's α between 0.80 and 0.90 in several German samples [42]. Indications for validity is given through associations of the GSE score with other psychological constructs, e.g. negative correlations with depression, anxiety, and burnout.

Moreover, self-perceived health status was rated using five categories ("Very bad", "Bad", "Moderate", "Good", and "Very good") [43]. Demographic data, for example, sex, age, education and working status were assessed as well as net income. This was done in five categories, matching the income of the elderly in Austria (www.statistik.at).

Sample size

Sample size considerations were based on detecting differential item functioning (DIF) in another questionnaire used in this study, the Patient Activation Measure® [44]. DIF assesses whether an item measures the same abilities across subgroups. There is a suggestion to estimate the power of a Rasch model via DIF [45]. Simulations indicate that power increases with larger sample sizes and fewer items, improving DIF detection. For example, with a 20-item questionnaire, power rises from 43% with 50 participants to 79% with 100 participants. While no simulations with larger samples were conducted, it's evident that power improves with sample size. Regarding the Patient Activation Measure®, DIF was found for some items in a large sample of 4300 participants [46]. Given the Patient Activation's structure (13 items with 4 response options), a sample of at least 500 is recommended. To ensure adequate power to detect potential DIF, the goal was to interview N = 700 patients.

Data analysis

Categorical data are displayed using absolute and relative frequencies, while continuous data are represented by means and standard deviations, or alternatively, medians and interquartile ranges, as suitable.

The framework of IRT was used to analyze DIF in the GSE. The assumption of IRT analysis of unidimensionality was examined through a confirmatory factor analysis (CFA). After model comparison using fit-indices, LR-test and Vuong tests, the graded response model (GRM) was used to represent the data. After estimating the GRM with all available data, missing responses ($N = 15$, 0.3% of all values) were imputed based on the latent abilities estimated by the model. To account for uncertainty in the imputation process, multiple imputations (1000 times) for missing responses were performed, and each missing value was replaced by the rounded average value of its imputations. Root mean square error of approximation (RMSEA), standardized root mean square residual (SRMSR), Tucker-Lewis index (TLI) and comparative fit index (CFI) were used to assess model fit. A good fit was defined as a RMSEA < 0.05, SRMSR ≤ 0.08 and > 0.9 for TLI and CFI [47]. Moreover, sample adjusted Bayesian information criterion (SABIC) was examined, smaller values indicate a better model fit. To investigate model fit at the item level, infit and outfit statistics were used. The range from 0.5 to 1.5 is typically suitable for measurement, while the range between -1.9 and 1.9 indicates reasonable predictability. Values ≤ -2 indicate highly predictable data [48].

To assess if there is a bias in the GSE, DIF was used. Characteristics of interest were interviewer (MH/VW/AK/JG), sex (male/female), age (median split, < 70 / ≥ 70), education (no degree/degree), working status (retired/other), income (< 1,125 / ≥ 1,125—≤ 1,950/ ≥ 1,950), type of macular edema (diabetic, retinal vein occlusion or both), diagnoses (Hypertension, diabetes, or

both), and health status (very good – good/moderate/bad – very bad). To detect DIF, the likelihood-ratio $\chi 2$ test was used. For assessing DIF magnitude McFadden's pseudo-R^2 and non-compensatory differential item functioning (NCDIF) were used. The latter is a statistic on the item level reflecting differences in the scores of two distinct characteristics. The kind of DIF, uniform or non-uniform was investigated as well. Moreover, differential test functioning (DTF) was investigated. DTF describes the accumulation of DIF effects of all items on the whole questionnaire, which means the overall score is dependent on a characteristic and different for distinct groups, e.g. males and females. To further investigate potential DIF induced bias, Bland–Altman analyses were conducted comparing the original model and the biased model, in case DIF was found for a characteristic. For estimating IRT models, the package mirt [49] and for estimating DIF the package lordif [50] were used, using R studio version 4.1.1 [51].

Results

Study participants

Responses of $N=556$ patients were available for data analysis. Of all four interviewers, MH conducted 187 (34%), VW 159 (29%), AK 114 (21%) and JG 96 (17%) interviews, where each interviewer performed as many interviews as feasible. The GSE was finished in about 1.7 min (IQR: 1.5 – 2.2).

The median age was 68.4 (IQR: 62.0 – 76.0) and $N=322$ (58%) of patients were male. A macular edema due to diabetes was present in 319 (57%) patients, due to retinal vein occlusion in 224 (40%) and 13 (2%) exhibited both types, for further characteristics see Table 1.

Table 1 Socio-demographic and health characteristics ($n=556$)

	Total
Interviewer	
MH	187 (34%)
VW	159 (29%)
AK	114 (21%)
JG	96 (17%)
Age in years	68.4 (62.0 – 76.0)
< 70 years	279 (50%)
≥ 70 years	277 (50%)
Sex	
Male	322 (58%)
Female	234 (42%)
Education	
No degree: basic education	413 (74%)
Degree (High school or higher Education)	143 (26%)
Working status	
Other[a]	107 (19%)
Retired	449 (81%)
Monthly net income	
< 800€	73 (13%)
< 1125€	92 (17%)
< 1500€	112 (21%)
≤ 1950€	82 (15%)
> 1950€	184 (34%)
Missing	13 (2%)
BMI[b]	26.8 (24.7 – 30.0)
Type of macular edema	
Diabetic	319 (57%)
Retinal vein occlusion	224 (40%)
Both	13 (2%)
Diagnosis	
Diabetes	129 (23%)
Hypertension	196 (35%)
Both	231 (42%)
Health status self-rated	
Very good	61 (11%)
Good	236 (43%)
Moderate	223 (40%)
Bad	30 (5%)
Very bad	6 (1%)

Data are presented as N (%), median (25th – 75th percentiles)

[a] Category 'Other' includes working ($N=66$, 12%), jobless, studying and homemaker

[b] *BMI* Body mass index

GSE results

The most frequent response category across all items was "Agree", which was selected between 43% (item 3) and 58% (item 10) of patients. "Disagree strongly" was the least chosen response category with 1% (item 7) to 4% (item 4). Mean response scores ranged from 3.1 (SD 0.7) for item 10 to 3.4 (SD 0.6) for item 7, see Table 2. Participants showed a mean GSE score of 32.8 (SD: 4.8). The mean GSE score for patients interviewed by MH was 32.6 (SD: 4.8), by VW 32.4 (SD: 5.2), by AK 32.6 (SD: 4.5) and by. JG 33.5 (SD: 4.6). Mean GSE scores for other relevant characteristics see supplementary Table 1.

Psychometric properties

First, the IRT model was estimated with the initial four categories. However, due to the low frequencies of responses for the lowest category, there were problems in estimating the DIF. Therefore, the two lowest response categories were collapsed, a procedure that is statistically justifiable as collapsing ordered response categories is permissible due to the artificial nature of the category divisions [52], and previous research has shown that parameter estimates and theta estimates remain similar and highly correlated despite such collapses [53, 54].

Table 2 Interviewer effect on the items of the general self-efficacy scale

					NCDIF		
Nr	Item	β	R^2	p	MH vs. VW	MH vs. AK	MH vs. JG
1	I can always manage to solve difficult problems if I try hard enough	0.004	0.006	0.413	0.008	0.015	0.015
2	If someone opposes me, I can find the means and ways to get what I want	0.005	0.006	0.604	0.003	0.005	0.000
3	It is easy for me to stick to my aims and accomplish my goals	0.005	0.002	0.865	0.008	0.001	0.007
4	I am confident that I could deal efficiently with unexpected events	0.003	0.004	0.630	0.011	0.011	0.007
5	Thanks to my resourcefulness, I know how to handle unforeseen situations	0.002	0.007	0.522	0.003	0.006	0.008
6	I can solve most problems if I invest the necessary effort	0.004	0.002	0.939	0.000	0.004	0.000
7	I can remain calm when facing difficulties because I can rely on my coping abilities	0.014	0.006	0.628	0.005	0.006	0.010
8	When I am confronted with a problem, I can usually find several solutions	0.000	0.006	0.336	0.009	0.005	0.010
9	If I am in trouble, I can usually think of a solution	0.010	0.008	0.458	0.002	0.004	0.005
10	I can usually handle whatever comes my way	0.001	0.005	0.572	0.003	0.009	0.000

DTF: reference interviewer MH, VW: 0.010, AK: 0.007, JG: 0.024

β: influence of interviewers, R^2: pseudo-R^2 by McFadden, p values derived from a likelihood ratio $\chi 2$ test

GSE General Self-Efficacy Scale, *DTF* differential test functioning, *NCDIF* Non-compensatory differential item functioning

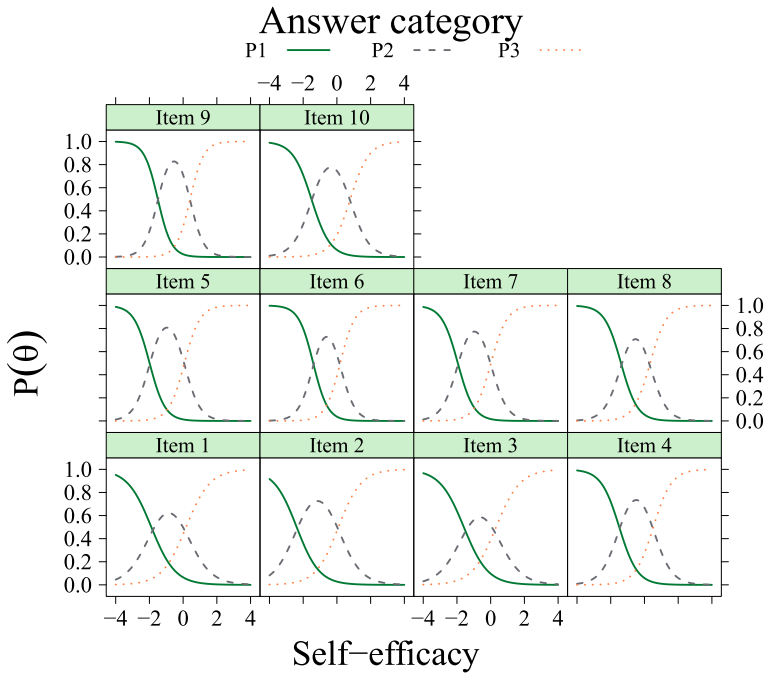

Fig. 1 Item characteristic curves for the General Self-Efficacy Scale estimated by a graded response model. P1: Not at all true & Barley true P2: Moderately true P3: Exactly true

The unidimensionality of the GSE scale was confirmed by a CFA, and 55% of the variance was explained by one factor.

The final GRM showed a RMSEA of 0.08 (90% CI: 0.07 – 0.09), SRMSR of 0.05, TLI of 0.96 and a CFI of 0.97, indicating a good model fit. In general items showed good in- and outfit, except items 4, 5, and 6 showing slightly bad outfit and item 9 slightly bad infit. Item 2 showed the lowest difficulty of -1.15 and item 10 showed the highest difficulty with -0.36, see Fig. 1. Item discrimination ranged from 1.40 for item 3 to 2.47 for item 9, see supplementary Table 2. The GRM estimated an empirical reliability of 0.86, see Fig. 2.

Differential item functioning

No DIF was observed with respect to the grouping variable interviewer. In other words, there was no evidence to suggest that respondents' item responses systematically

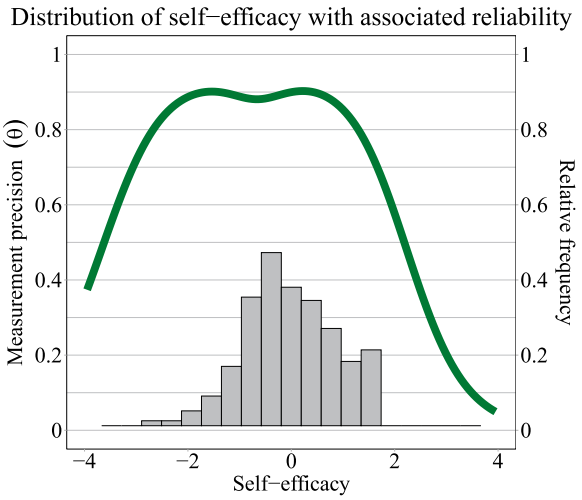

Fig. 2 Estimated reliability of the General Self-Efficacy Scale estimated by a graded response model (green line) and the distribution of ability in the sample (grey histogram). The x-axis represents the person ability (in logits), while the y-axis depicts measurement precision (0 – 1) and relative frequency of person abilities (0 – 1)

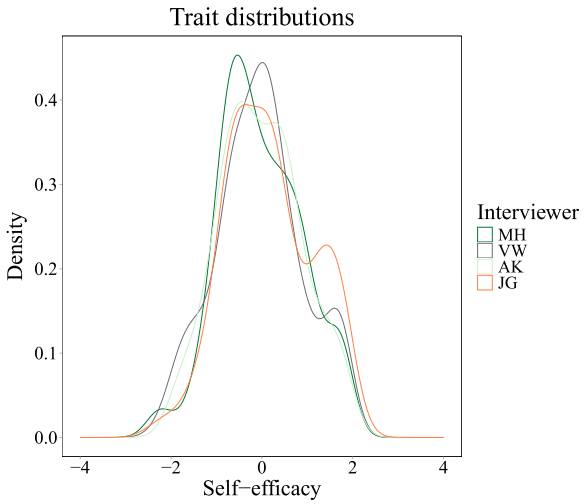

Fig. 3 Differential test functioning for interviewer bias for the General Self-Efficacy Scale over the ability range in logits

varied based on the specific interviewer who administered the questionnaire, see Table 2 and Fig. 3. Moreover, no DIF was found for age, health status, type of macular edema and diagnoses. However, DIF was found for sex in item 3 ("It is easy for me to stick to my aims and accomplish goals."), see Table 3. In the very low ability of self-efficacy this item was easier to endorse for women, from medium low ability on it was easier to endorse for men. Moreover, DIF was detected for education in item 1 ("If someone opposes me, I can find means and ways to get what I want."). It was easier to endorse for patients with a degree compared to no degree. DIF was also found for working status in item 2 ("I can always manage to solve difficult problems if I try hard enough."), it was easier for patients if they were not retired. Furthermore, for income DIF was detected in item 1 and 3. Item 1 was mostly easier to endorse for patients with medium and high income compared to lower income, but in the high ability range, this item was easier to endorse with lower income than with medium income. Item 3 was easiest to endorse for patients with highest income and hardest for patients with lowest income.

Bland–Altman analysis was performed to compare the original (unbiased) model with the biased models due to education, income, working status and sex. No significant bias was fond, see supplementary Table 3, supplementary Fig. 1, supplementary Fig. 2, supplementary Fig. 3 and supplementary Fig. 4.

Discussion

This study investigated the influence of interviewers on the GSE responses in a face-to-face interview. No interviewer-related DIF was found, suggesting that objective assessment is possible with the GSE in an interview setting, if interviewers are properly trained. Another finding is that the GSE in the interviewer setting is measurement fair for different age, health status, type of macular edema and diagnosis groups. However, DIF was found for sex, income, education, and working status in some items.

Further indicating the feasibility of the interview setting is that the GSE score found in this study is similar to the norm values of the German version (32.8 SD: 4.81. vs. 29.4 SD: 5.36) [42]. The estimated reliability (0.86) was within the range as found in several German samples (between 0.80 and 0.90) [42]. Moreover, the scale was found to be unidimensional as in previous studies [14, 41, 42] and was best described with a graded response model [55].

Whenever the influence of interviewers on the response is systematical, a response bias occurs. One important cause for this bias is the presence of an interviewer, which may lead respondents to provide more socially acceptable answers [31]. Individuals may be inclined to present themselves in a more socially desirable light, leading to an overestimation of their capabilities and performance [31]. The concept of social desirability is strongly influenced by social norms and standards, as an individual's knowledge of norms and standards determines what they perceive as socially desirable or undesirable [56]. Social norms lead people to suppress critical feedback and hide their negative emotions, leading to overconfident self-impressions and inaccurate self-assessments [57].

Table 3 Differential item functioning found in the general self-efficacy scale

Item	Characteristic	Kind of DIF	β	R^2	p	Group	NCDIF	DTF
1	Education	Uniform	0.058	0.012	< 0.001		0.042	0.051
	Income	Non-uniform	0.042	0.009	0.007	1 vs. 2	0.050	0.127
						1 vs. 3	0.099	0.370
2	Working status	Uniform	0.013	0.011	0.004		0.017	0.033
3	Sex	Non-uniform	0.028	0.007	0.005		0.043	0.037
	Income	Uniform	0.041	0.022	< 0.001	1 vs. 2	0.024	0.127
						1 vs. 3	0.095	0.370

B influence of interviewers, R^2 pseudo-R^2 by McFadden, *p* values derived from a likelihood ratio χ2 test. *GSE* General Self-Efficacy Scale, *DIF* differential item functioning, *DTF* differential test functioning, *NCDIF* Non-compensatory differential item functioning

Furthermore, social norms and self-efficacy influence each other. Self-efficacy influences the perception and internalization of social norms and behaviors [58]. When an important social group sets a standard for behavior (social norm), people behaving similarly feel more self-efficacious [59]. Moreover, self-efficacy ratings and social desirability are connected [60, 61].

Another possible source of interviewer bias is the effect of interviewer characteristics. These include the interviewer's tone of voice, body language, and level of engagement with the respondent [28]. In addition, the interviewer's own biases, subjectivities, and lack of interviewing skills can introduce additional distortions into data collection. This can result in inaccurate or unreliable measurements [28]. This study did not find a possible interviewer bias introduced by the four different interviewers conducting the interviews. This indicates a consistent pattern of responses across different interviewers, supporting the objectivity of the questionnaire results.

To the best of our knowledge, interviewer bias has not been assessed in the GSE. Furthermore, the use of DIF as a method to detect interviewer bias is rare [32–34, 62]. In other studies, questionnaires were administered through interviews, but DIF was only examined concerning other characteristics (e.g. [62, 63]). Other studies have investigated interviewer effects using a different method, namely hierarchical regression models [64–68]. One of these studies showed that interviewers have a significant impact on responses, with interviewer effects even outweighing respondent differences [67]. Moreover, interview bias occurred before in different fields. For example, in social and political issues, participants adjusted their answers according to their perception of the interviewer [30]; interviewer bias was found in clinical interviews diagnosing borderline personality disorder [62] and in a European social survey [67]. In another study, interviewer bias was induced by the sex of the interviewer [69]. Even though there was no interviewer bias present in this study, these findings underscore the importance of investigating the effect of an interviewer on participants' answers.

In this study, more characteristics regarding DIF were investigated in an interview setting. Only a few items (items 1, 2, and 3) raised concerns. Comparable to our study, DIF for education on the GSE was found before, but other items were affected [24]. One possible explanation is that individuals with higher self-efficacy show better academic outcomes and are more likely to finish a degree [7, 70–72]. On the one hand, self-efficacy predicts the performance of educational requirements [70], and on the other hand, the amount of experiences is related to students' self-efficacy [72]. It may be that solving more complex problems during advanced education leads degree holders with similar self-efficacy to agree more readily with an item, due to the experience gained while completing their degree, as increased experience is positively associated with self-efficacy [73].

In addition, DIF was found for two items related to income, a characteristic not previously investigated in this context. The items were easier to endorse for participants with higher incomes. One possible explanation is that having sufficient financial resources may help to cope with problems related to illness. Further, individuals with lower incomes are more likely to associate everyday situations with financial concerns, as they tend to notice the cost implications more quickly. These economic anxieties are natural and are difficult to ignore and affect the way they perceive and relate to different aspects of their lives [74].

Working status had an impact on participants' response patterns, which has been reported before [24]. Usually, workers show more self-efficacy than non-workers, including retirees and unemployed individuals [75]. A possible explanation for DIF found in this study is that self-efficacy decreases with higher age [7, 14], and this sample consisted mainly of retirees.

Moreover, DIF was found for sex. While Lönnfjord & Hagquist, 2017 also found DIF for the same item, others

found DIF for other items [19, 20] or did not find any DIF [21–23]. In general, men tend to rate their self-efficacy higher than women [24, 42, 75, 76]. Sex differences are also observed in other domains, such as academic self-efficacy, which varies depending on the subject area [77]. Specifically, women demonstrated the highest level of academic self-efficacy in language arts, whereas men exhibited the highest levels in mathematics, computer science, and social science. Furthermore, both sexes rate their self-efficacy different, depending on the complexity of the task [78]. Specifically, men exhibit higher self-efficacy for more complex tasks. It is likely that in this study, women found it easier to endorse items in the very low ability range, while men found it easier to endorse items in the medium to low ability level.

Overall, DIF was detected for some characteristics in some items, its impact on the overall results appears negligible: differences in scores due to bias were found to be. Under conditions of adequate interviewer training, no additional procedures to correct the GSE score need to be implemented in clinical applications.

Limitations

The current study had potential limitations. These findings are specific to the German GSE read aloud and may not be generalizable to other interview situations, such as those involving non-standardized questionnaires or more free-form conversations. Since this study was not intended to examine the potential effects associated with certain characteristics of interviewers, interviewers were selected due to other criteria and all interviewers were women. One known possible influence on participants' answers is the sex of the interviewer [79–81].Participants may alter their responses based on the sex of the interviewer, possibly influenced by social desirability biases [79]. The perceived similarity between interviewer and respondent may moderate the relationship between social norms and self-efficacy [82]. Therefore, a limitation of this study is the absence of male interviewers. Incorporating male interviewers administering the GSE orally would have provided valuable insights into whether the face-to-face setting is feasible across more diverse interview settings, providing the possibility to investigate whether perceived similarity based on sex is to be considered.

Additionally, addressing sensitive topics in interviews presents a challenge, potentially affecting the honesty and accuracy of participant responses. Misreporting on sensitive survey topics is common and largely situational, with participants avoiding embarrassment or consequences from interviewers [83]. In this study, efforts were made to minimize this potential bias by fostering a welcoming atmosphere characterized by non-judgmental interactions and the clarification that there are no wrong responses.

One further limitation of the current study is that it did not explicitly examine the potential impact of nonverbal cues, such as tone, body language, and interpersonal dynamics, on respondent answers in face-to-face interviews, which may have influenced the results.

The study included patients from diverse socioeconomic backgrounds, which enhances the generalizability of the findings to some extent. However, the influence of diverse cultural backgrounds on the results is unclear and requires further investigation. While the generalizability of our findings may be limited, we assume that the results are likely applicable to populations with chronic conditions within a comparable age range. Specifically, our findings may be particularly relevant to individuals managing conditions such as hypertension and diabetes mellitus, since most of our study sample suffered from at least one of those diseases.

Conclusions

Overall, the results of this study indicate that the German version of the GSE can be used in a face-to-face interview with chronic disease patients without compromising its psychometric properties. If interviewers are properly trained, they may not strongly influence the answers given by participants. Due to the impact of the interviewer setting on patients, it is crucial to implement a standardized procedure to ensure objective assessment [28]. Some issues regarding DIF concerning sex, education, working status and income occurred. Since these DIF effects found were small, it appears that the GSE is a fair measurement instrument for assessing self-efficacy in an interview setting. One application of the GSE is to identify individuals who might benefit from self-efficacy enhancement interventions in clinical settings. The GSE scale has the potential to be made accessible to a broader population, including individuals with visual impairment, as it may be the case that self-efficacy is of interest with individuals that are unable to complete the GSE by themselves. For example, diabetic patients with reading difficulties due to visual impairment, often have problems with their eyes, leading to visual impairment. Increasing their self-efficacy can promote better health outcome [84]. As further research, investigating the feasibility of training healthcare staff to administer the GSE in an interview setting, while ensuring that the answers given are bias-free, would be valuable.

Abbreviations

BMI	Body mass index
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CTT	Classical test theory

DIF	Differential item functioning
DTF	Differential test functioning
GSE	General self-efficacy scale
GRM	Graded response model
IRT	Item response theory
NCDIF	Non-compensatory differential item functioning
RMSEA	Root mean square error of approximation
SABIC	Sample adjusted Bayesian information criterion
SRMSR	Standardized root mean square residual
TLI	Tucker-Lewis index

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40359-025-02579-2.

Supplementary Material 1: Supplementary table 1. GSE score for relevant characteristics. Supplementary table 2. GSE item statistics (n = 556). Supplementary table 3. Bland-Altman statistics. Supplementary figure 1. Bland-Altman plot comparing the General Self-Efficacy Scale (GSE) estimated by a graded response model with and without bias for education. The x-axis represents the mean of the two GSE scores for each participant, while the y-axis shows the difference between the original and biased scores (original minus biased). The solid horizontal line indicates the mean difference (bias) between the two methods. The dashed lines represent the 95% limits of agreement. These limits suggest that for 95% of cases, the differences between adjusted and unadjusted GSE scores are expected to fall within this range. Supplementary figure 2. Bland-Altman plot comparing the General Self-Efficacy Scale (GSE) estimated by a graded response model with and without bias for income. The x-axis represents the mean of the two GSE scores for each participant, while the y-axis shows the difference between the original and biased scores (original minus biased). The solid horizontal line indicates the mean difference (bias) between the two methods. The dashed lines represent the 95% limits of agreement. These limits suggest that for 95% of cases, the differences between adjusted and unadjusted GSE scores are expected to fall within this range. Supplementary figure 3. Bland-Altman plot comparing the General Self-Efficacy Scale (GSE) estimated by a graded response model with and without bias for sex. The x-axis represents the mean of the two GSE scores for each participant, while the y-axis shows the difference between the original and biased scores (original minus biased). The solid horizontal line indicates the mean difference (bias) between the two methods. The dashed lines represent the 95% limits of agreement. These limits suggest that for 95% of cases, the differences between adjusted and unadjusted GSE scores are expected to fall within this range. Supplementary figure 4. Bland-Altman plot comparing the General Self-Efficacy Scale (GSE) estimated by a graded response model with and without bias for working status. The x-axis represents the mean of the two GSE scores for each participant, while the y-axis shows the difference between the original and biased scores (original minus biased). The solid horizontal line indicates the mean difference (bias) between the two methods. The dashed lines represent the 95% limits of agreement. These limits suggest that for 95% of cases, the differences between adjusted and unadjusted GSE scores are expected to fall within this range.

Acknowledgements
We thank all the staff members at the Department of Ophthalmology for making this study possible and the patients for their cooperation.

Authors' contributions
MH, AA, AB and AW took part in planning the study; MW, SS and MM, TF, MGA, MGR recruited patients; MH, VW and AK did data assessment; MH and AA data analysis; MH main part writing manuscript; all authors contributed to the manuscript.

Authors' information
We confirm that all authors were involved in the design and implementation of the study, or acquisition of data, or analysis and interpretation of data. All authors contributed and approved to the manuscript.

Funding
This manuscript represents a part of a project that did not receive any specific grant from funding agencies in the public, commercial, or non-profit sectors and was sponsored by the Medical University of Graz.

Data availability
The participants of this study did not give written consent for their data to be shared publicly, so due to the sensitive nature of the research the data cannot be made available. Therefore, data is provided within the manuscript and supplementary files.

Declarations

Ethics approval and consent to participate
The ethical committee of the Medical University of Graz approved the study (32–101 ex 19/20) in accordance with the Declaration of Helsinki. Participants provided written informed consent to participate in this study and were able to opt out at any moment.

Consent for publication
Not applicable. All personal data are deidentified.

Competing interests
The authors declare no competing interests.

Received: 3 July 2024 Accepted: 6 March 2025
Published online: 25 March 2025

References

1. Zimmerman. Self-Efficacy: an essential motive to learn. Contemp Educ Psychol. 2000;25 1:82–91.
2. Kadden RM, Litt MD. The role of self-efficacy in the treatment of substance use disorders. Addict Behav. 2011;36(12):1120–6.
3. Bandura A. Perceived self-efficacy in cognitive development and functioning. Educ Psychol. 1993;28(2):117–48.
4. Banik A, Schwarzer R, Knoll N, Czekierda K, Luszczynska A. Self-efficacy and quality of life among people with cardiovascular diseases: a meta-analysis. Rehabil Psychol. 2018;63(2):295–312.
5. Jones F, Riazi A. Self-efficacy and self-management after stroke: a systematic review. Disabil Rehabil. 2011;33(10):797–810.
6. Patricio-Gamboa R, Alanya-Beltrán J, Acuña-Condori SP, Poma-Santivañez Y. Perceived self-efficacy geared towards education: systematic review. Espirales Rev Multidiscip Investig. 2021;5(37):32–45.
7. Luszczynska A, Gutiérrez-Doña B, Schwarzer R. General self-efficacy in various domains of human functioning: Evidence from five countries. Int J Psychol. 2005;1(40):80–9.
8. Scherbaum CA, Cohen-Charash Y, Kern MJ. Measuring general self-efficacy: a comparison of three measures using item response theory. Educ Psychol Meas. 2006;66(6):1047–63.
9. Wahyuni A, Ramayani D. The relationship between self-eficacy and self-care in type 2 diabetes mellitus patients. Malays J Nurs. 2020;11(03):68–75.
10. Picha KJ, Howell DM. A model to increase rehabilitation adherence to home exercise programmes in patients with varying levels of self-efficacy. Musculoskeletal Care. 2018;16(1):233–7.
11. Tilden EL, Caughey AB, Lee CS, Emeis C. The effect of childbirth self-efficacy on perinatal outcomes. J Obstet Gynecol Neonatal Nurs. 2016;45(4):465–80.
12. Miles C, Pincus T, Carnes D, Taylor S, Underwood M. Measuring pain self-efficacy. Clin J Pain. 2011;1(27):461–70.
13. Jacob ME, Lo-Ciganic WH, Simkin-Silverman LR, Albert SM, Newman AB, Terhorst L, et al. The preventive services use self-efficacy (PRESS) scale in older women: development and psychometric properties. BMC Health Serv Res. 2016;16(1):71.
14. Schwarzer R, Jerusalem M. The general self-efficacy scale (GSE). Anxiety Stress Coping. 2010;12(1):329–45.

15. Lee C, Bobko P. Self-efficacy beliefs: Comparison of five measures. J Appl Psychol. 1994;79(3):364–9.
16. Burrell AMG, Allan JL, Williams DM, Johnston M. What do self-efficacy items measure? Examining the discriminant content validity of self-efficacy items. Br J Health Psychol. 2018;23(3):597–611.
17. Bandura A. Guide for constructing self-efficacy scales. Self-Effic Beliefs Adolesc. 2006;5(1):307–37.
18. Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. Patient - Patient-Centered Outcomes Res. 2014;7(1):23–35.
19. Damasio B, Valentini F, Nunez S, Kliem S, Koller S, Hinz A, et al. Is the general self-efficacy scale a reliable measure to be used in cross-cultural studies? Results from Brazil, Germany and Colombia. Span J Psychol. 2016;26:19.
20. Lönnfjord V, Hagquist C. The psychometric properties of the Swedish version of the general self-efficacy scale: a rasch analysis based on adolescent data. Curr Psychol N B Nj. 2017;37:703–15.
21. Peter C, Cieza A, Geyh S. Rasch analysis of the general self-efficacy scale in spinal cord injury. J Health Psychol. 2014;19(4):544–55.
22. Salsman JM, Schalet BD, Merluzzi TV, Park CL, Hahn EA, Snyder MA, et al. Calibration and initial validation of a general self-efficacy item bank and short form for the NIH PROMIS®. Qual Life Res Int J Qual Life Asp Treat Care Rehabil. 2019;28(9):2513–23.
23. Sun V, Raz DJ, Ruel N, Chang W, Erhunmwunsee L, Reckamp K, et al. A multimedia self-management intervention to prepare cancer patients and family caregivers for lung surgery and postoperative recovery. Clin Lung Cancer. 2017;18(3):e151–9.
24. Bonsaksen T, Kottorp A, Gay C, Fagermoen MS, Lerdal A. Rasch analysis of the general self-efficacy scale in a sample of persons with morbid obesity. Health Qual Life Outcomes. 2013;11(1):202.
25. Eslami A, Daniali SS, Mohammadi K, Reisi-Dehkordi N, Mostafavi-Darani F. Cultural adaptation and psychometric properties of the persian version of self-efficacy in chronic disease patients. Iran J Nurs Midwifery Res. 2017;22(1):57–61.
26. Leeuw ED de, Hox JJ, Dillman DA, European Association of Methodology, editors. International handbook of survey methodology. New York ; London: Lawrence Erlbaum Associates; 2008. 549 p. (EAM book series).
27. Cook C. Mode of administration bias. J Man Manip Ther. 2010;18(2):61–3.
28. Hofisi C, Hofisi M, Mago S. Critiquing interviewing as a data collection method. Mediterr J Soc Sci. 2014;1:5.
29. Barath A, Cannell CF. Effect of Interviewer's Voice Intonation. Public Opin Q. 1976;40(3):370–3.
30. Kühne S. Interpersonal perceptions and interviewer effects in face-to-face surveys. Sociol Methods Res. 2023;52(1):299–334.
31. Bowling A. Mode of questionnaire administration can have serious effects on data quality. J Public Health. 2005;27(3):281–91.
32. Kisala PA, Boulton AJ, Cohen ML, Slavin MD, Jette AM, Charlifue S, et al. Interviewer- vs self-administration of PROMIS measures for adults with traumatic injury. Health Psychol Off J Div Health Psychol Am Psychol Assoc. 2019;38(5):435.
33. Rausch-Koster TP, Luijten MAJ, Verbraak FD, van Rens GHMB, van Nispen RMA. Calibration of the dutch eyeQ to measure vision related quality of life in patients with exudative retinal diseases. Transl Vis Sci Technol. 2022;11(4):5.
34. Holter M, Avian A, Weger M, Strini S, Michelitsch M, Brenk-Franz K, et al. Measuring patient activation: the utility of the Patient Activation Measure administered in an interview setting. Qual Life Res. 2024 Feb 22; https://doi.org/10.1007/s11136-024-03614-2. Cited 2024 Mar 6.
35. Broering JM, Paciorek A, Carroll PR, Wilson LS, Litwin MS, Miaskowski C. Measurement equivalence using a mixed-mode approach to administer health-related quality of life instruments. Qual Life Res. 2014;23(2):495–508.
36. Sikorskii A, Noble PC. Statistical considerations in the psychometric validation of outcome measures. Clin Orthop Relat Res. 2013;471(11):3489–95.
37. Sim J ah, Hyun G, Gibson TM, Yasui Y, Leisenring W, Hudson MM, et al. Negligible Effects of the Survey Modes for Patient-Reported Outcomes: A Report From the Childhood Cancer Survivor Study. JCO Clin Cancer Inform. 2020 Jan 17; Available from: https://ascopubs.org/doi/https://doi.org/10.1200/CCI.19.00135. Cited 2024 Mar 13.
38. Spangenberg L, Glaesmer H, Boecker M, Forkmann T. Differences in patient health questionnaire and aachen depression item bank scores between tablet versus paper-and-pencil administration. Qual Life Res. 2015;24(12):3023–32.
39. Institute of Medical Informatics, Statistics and Documentation, Medical University of Graz, Austria. 2023. Available from: https://imi.medunigraz.at/en/services#c44617. Cited 2022 Sep 15.
40. LimeSurvey - einfache Online-Umfragen. 2023. Available from: https://www.limesurvey.org/de/. Cited 2022 Sep 11.
41. Schwarzer R, Jerusalem M. Skalen zur erfassung von Lehrer-und schülermerkmalen. In Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) (Hrsg.), Elektronisches Testarchiv. Trier: ZPID.; 1999. Available from: https://www.testarchiv.eu/index.php?wahl=testarchiveintro.
42. Hinz A, Schumacher J, Albani C, Schmid G, Brähler E. Bevölkerungsrepräsentative Normierung der Skala zur Allgemeinen Selbstwirksamkeitserwartung. Diagnostica. 2006;52(1):26–32.
43. De Bruin A. Health Interview Surveys: Towards International Harmonization of Methods and Instruments. WHO Regional Publications, European Series, No. 58. ERIC; 1996.
44. Hibbard JH, Mahoney ER, Stockard J, Tusler M. Development and testing of a short form of the patient activation measure. Health Serv Res. 2005;40(6p1):1918–30.
45. Kubinger KD, Rasch D, Yanagida T. On designing data-sampling for Rasch model calibrating an achievement test. Psychol Test Assess Model. 2009;51(4):370.
46. Zill JM, Dwinger S, Kriston L, Rohenkohl A, Harter M, Dirmaier J. Psychometric evaluation of the German version of the Patient Activation Measure (PAM13). BMC Public Health. 2013;13(Journal Article):1027–2458-13-1027.
47. Hu L tze, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model Multidiscip J. 1999;6(1):1–55.
48. Linacre JM. What do Infit and Outfit, mean-square and standardized mean? Rasch Meas Trans. 2002;16:878.
49. Chalmers RP. mirt: a multidimensional item response theory package for the R environment. J Stat Softw. 2012;48:1–29.
50. Choi SW, Crane PK, Choi MSW. Package 'lordif'. 2016. Available from: http://r.meteo.uni.wroc.pl/web/packages/lordif/lordif.pdf. Cited 2024 Jun 5.
51. RStudio RsT. Integrated development environment for R. RStudio PBC Boston MA USA. 2020.
52. Jansen PGW, Roskam EE. Latent trait models and dichotomization of graded responses. Psychometrika. 1986;51(1):69–91.
53. Jeong HJ, Lee WC. The level of collapse we are allowed: comparison of different response scales in safety attitudes questionnaire. Biom Biostat Int J. 2016;4(4):00100.
54. Muraki E. Fitting a polytomous item response model to likert-type data. Appl Psychol Meas. 1990;14(1):59–71.
55. Sun X, Zhong F, Xin T, Kang C. Item response theory analysis of general self-efficacy scale for senior elementary school students in China. Curr Psychol. 2021;40(2):601–10.
56. BouMalham P, Saucier G. The conceptual link between social desirability and cultural normativity: desirability and Normativity. Int J Psychol. 2016;51(6):474–80.
57. Fay AJ, Jordan A, Ehrlinger J. How social norms promote misleading social feedback and inaccurate self-assessment. Soc Personal Psychol Compass. 2012;6:206–16.
58. Chung A, Rimal RN. Social norms: a review. Rev Commun Res. 2016;4:1–28.
59. Stok FM, Verkooijen KT, de Ridder DTD, de Wit JBF, de Vet E. How norms work: self-identification, attitude, and self-efficacy mediate the relation between descriptive social norms and vegetable intake. Appl Psychol Health Well-Being. 2014;6(2):230–50.
60. Silverthorn NA, Gekoski WL. Social desirability effects on measures of adjustment to university, independence from parents, and self-efficacy. J Clin Psychol. 1995;51(2):244–51.
61. Jago R, Baranowski T, Baranowski JC, Cullen KW, Thompson DI. Social desirability is associated with some physical activity, psychosocial

variables and sedentary behavior but not self-reported physical activity among adolescent males. Health Educ Res. 2006;22(3):438–49.
62. Sharp C, Steinberg L, Michonski J, Kalpakci A, Fowler C, Frueh BC, et al. *DSM* borderline criterion function across age-groups: a cross-sectional mixed-method study. Assessment. 2019;26(6):1014–29.
63. Teresi JA, Ocepek-Welikson K, Kleinman M, Cook KF, Crane PK, Gibbons LE, et al. Evaluating measurement equivalence using the item response theory log-likelihood ratio (IRTLR) method to assess differential item functioning (DIF): applications (with illustrations) to measures of physical functioning ability and general distress. Qual Life Res. 2007;16(S1):43–68.
64. Hox J. Hierarchical regression models for interviewer and respondent effects. Sociol Methods Res. 1994;22:300–18.
65. Herzing JM, Blom AG, Meuleman B. Modeling group-specific interviewer effects on survey participation using separate coding for random slopes in multilevel models. J Surv Stat Methodol. 2024;12(1):249–73.
66. West BT, Blom AG. Explaining interviewer effects: a research synthesis. J Surv Stat Methodol. 2017;5(2):175–211.
67. Loosveldt G, Beullens K. Interviewer effects on non-differentiation and straightlining in the European social survey. J Off Stat. 2017;33(2):409–26.
68. Murphy J, Biemer P, Stringer C, Thissen R, Day O, Hsieh Y. Interviewer falsification: current and best practices for prevention, detection, and mitigation. Stat J IAOS. 2016;32:313–26.
69. Catania JA, Binson D, Canchola J, Pollack LM, Hauck W. Effects of interviewer gender, interviewer choice, and item wording on responses to questions concerning sexual behavior. Public Opin Q. 1996;60(3):345.
70. Schunk DH. Self-efficacy and achievement behaviors. Educ Psychol Rev. 1989;1(3):173–208.
71. Mangus L, Somers C, Yoon J, Partridge T, Pernice F. Examination of college student achievement within an ecological framework. J Adult Contin Educ. 2021;27(2):231–47.
72. van Dinther M, Dochy F, Segers M. Factors affecting students' self-efficacy in higher education. Educ Res Rev. 2011;6(2):95–108.
73. Usher EL, Pajares F. Sources of self-efficacy in school: critical review of the literature and future directions. Rev Educ Res. 2008;78(4):751–96.
74. Shah AK, Zhao J, Mullainathan S, Shafir E. Money in the mental lives of the poor. Soc Cogn. 2018;36(1):4–19.
75. Bonsaksen T, Lerdal A, Heir T, Ekeberg Ø, Skogstad L, Grimholt T, et al. General self-efficacy in the Norwegian population: differences and similarities between sociodemographic groups. Scand J Public Health. 2019;47:695–704.
76. Nielsen T, Dammeyer J, Vang ML, Makransky G. Gender fairness in self-efficacy? A rasch-based validity study of the General Academic Self-Efficacy Scale (GASE). Scand J Educ Res. 2018;62(5):664–81.
77. Huang FY, Chung H, Chung H, Chung H, Kroenke K, Delucchi KL, et al. Using the Patient Health Questionnaire-9 to Measure Depression among Racially and Ethnically Diverse Primary Care Patients. J Gen Intern Med. 2006;
78. Busch T. Gender differences in self-efficacy and attitudes toward computers. J Educ Comput Res. 1995;12(2):147–58.
79. Johnson TP, Braun M. Challenges of comparative survey research [Internet]. SAGE London; 2016. Available from: https://books.google.at/books?hl=de&lr=&id=g8OMDAAAQBAJ&oi=fnd&pg=PA41&dq=Johnson,+Timothy+P.,+and+Michael+Braun.+2016.+%E2%80%9CChallenges+of+Comparative+Survey+Research.%E2%80%9D+In+The+SAGE+Handbook+of+Survey+Methodology,+edited+by+Christof+Wolf,+Joye+Dominique,+Tom+W.+Smith,+and+Yang-chih+Fu,+41%E2%80%9354.+London:+SAGE.&ots=DAmHpyTZmX&sig=ruHppM65EqXOxPkJMhfytOKgkDM. Cited 2024 Apr 10.
80. Schmiedeberg C, Schröder J. Did you like the interview? Interviewer effects on respondents' interview pleasantness ratings. Field Methods. 2024;36(1):21–36.
81. Sundström A, Stockemer D. Measuring support for women's political leadership. Public Opin Q. 2022;86(3):668–96.
82. Rimal RN, Lapinski MK, Cook RJ, Real K. Moving toward a theory of normative influences: how perceived benefits and similarity moderate the impact of descriptive norms on behaviors. J Health Commun. 2005;10(5):433–50.
83. Tourangeau R, Yan T. Sensitive questions in surveys. Psychol Bull. 2007;133(5):859–83.
84. Jiang S, Wang C, Weiss DJ. Sample size requirements for estimation of item parameters in the multidimensional graded response model. Front Psychol. 2016;7(1):109.

Publisher's Note