The Journal of General Psychology

ISSN: 0022-1309 (Print) 1940-0888 (Online) Journal homepage: www.tandfonline.com/journals/vgen20

Age and gender differences in the factor structure of cognitive monitoring

Martin Komarc, Lawrence M. Scheier & Jana Novotná

To cite this article: Martin Komarc, Lawrence M. Scheier & Jana Novotná (23 Jun 2026): Age and gender differences in the factor structure of cognitive monitoring, The Journal of General Psychology, DOI: 10.1080/00221309.2026.2689111

To link to this article: https://doi.org/10.1080/00221309.2026.2689111

THE JOURNAL OF GENERAL PSYCHOLOGY
https://doi.org/10.1080/00221309.2026.2689111

Age and gender differences in the factor structure of cognitive monitoring

Martin Komarc[a], Lawrence M. Scheier[b], and Jana Novotná[a]

[a]Faculty of Physical Education and Sport, Charles University; [b]LARS Research Institute, Inc.

ABSTRACT

Cognitive development during adolescence is arguably unmatched by any other period of life. Enhanced brain maturation, increased memory capacity, and the acquisition of new cognitive skills help prepare adolescents for adult social roles. Among these skills is cognitive monitoring, or the ability to reflect on one's own thinking—traditionally referred to as "metacognition." Using cross-sectional data from the first wave of a longitudinal panel cohort, we examined three facets of cognitive monitoring: decision-making (e.g., gathering information, evaluating alternatives), self-reinforcement (e.g., praise and encouragement), and affective self-regulation (e.g., emotional control) in four age groups of Czech high school students. Multiple-group confirmatory factor analysis tested primary, higher-order, age and gender-based models. Overall, model fit was acceptable, with only trivial differences in item intercepts, factor correlations, variances, and latent factor means. Primary factor models revealed gender differences favoring girls' use of decision-making strategies and older youth employing more decision-making skills. A higher-order metacognitive factor to account for correlations among the monitoring skills was supported across both age and gender groups. Results are discussed in terms of promoting cognitive monitoring as a key 21st-century skill and advancing higher-order reflective skills as a critical learning task during adolescence.

ARTICLE HISTORY
Received 1 September 2025
Accepted 5 June 2026

KEYWORDS
invariance; adolescents; decision-making; self-reinforcement; affective self-management

Introduction

Children of all ages are notorious for talking to themselves while they engage in activities with mental reminders and encouragement. They use various internal self-talk strategies to remind themselves how to tackle a problem in school (e.g., "don't get frazzled, you know how to do this"), and push themselves during athletic competition (e.g., "if I run harder, I know that I can win this race"). This type of introspective verbal skill and

CONTACT Martin Komarc martin.komarc@ftvs.cuni.cz Department of Methodology, Faculty of Physical Education and Sport, Charles University, José Martího 31, Prague, Czech Republic

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00221309.2026.2689111.

cognitive self-guidance appears sometime in middle childhood in rudimentary form (e.g., Gascoine et al., 2017; Schneider et al., 2022). At some point, coinciding with adolescence, there is a growing awareness of the utility of cognitive monitoring and what Harris (1990) called "private speech" to the point where an individual recognizes when to use these strategies and the promise they yield (see also Brinthaupt, 2019 and Perrone-Bertolotti et al., 2014 for more on the "little voice in my head"). Flavell (1979) coined the term "metacognition" when he wrote about young children's inability to "monitor their own memory, comprehension, and other cognitive enterprises" suggesting that young children are "cognitive creatures with diverse cognitive tasks, goals, actions, and experiences" (p. 906). Recognition of the emergence of this type of monitoring was based on experimental work comparing younger to older age children on various cognitive and verbal interpretive tasks. Importantly, Flavell drew a wide arc around various literatures that share interest in the subject of metacognition including cognitive monitoring, control, and self-regulation, suggesting a new area of cognitive-developmental inquiry with ramifications for understanding the basis for more mature adultlike thought and behaviors (see also, Koriat, 2012).

Since Flavell's seminal writing on metacognition, considerable research has been conducted to further refine what is considered metacognition including its relations to learning (Veenman et al., 2006), intelligence (Veenman et al., 2005), motivation (Efklides, 2011), and critical thinking (Keating, 1990; Ku & Ho, 2010). In keeping with Flavell's original sentiments, Efklides (2008) suggested that metacognition is multidimensional and could be differentiated on the basis of knowledge, strategies, and experiences. At the heart of this distinction is knowing what is required of a task, how to approach it, the features of a task, whether it is hard, and different learning goals linked to the task (e.g., "this is going to be a hard test and I should study for it"). The actual strategies or cognitive procedures inherent in metacognition reflect the implementation schemes young children apply as they think about how to approach a problem (e.g., the features of a task, its demands, and how they will make progress toward their goals). Thus, metacognitive skills are akin to "procedural knowledge" in that they represent the conscious monitoring (i.e., introspection of what is required to complete the task) and strategic control (e.g., decision of what skill to apply) of cognition. The latter is what has made metacognition such a fuzzy construct and so difficult to distinguish from self-regulatory or control processes that involve an executive function (e.g., Dinsmore et al., 2008; Efklides, 2008; Roebers, 2017).

In the current study, we depart from the use of the term "metacognition" and rather use the term "cognitive monitoring" (e.g., Kuhn, 2000a; Schraw,

1998) to entail a type of self-awareness that one can apply and control mental strategies to accomplish the tasks, goals, actions, and experiences duly noted by Flavell. Strategically speaking, this entails planning, monitoring, and reflection (or evaluation, see, Schraw & Gutierrez, 2015) as principal skills and relies on other faculties (e.g., working memory, selective attention, error detection, and inhibitory control) to execute tasks (Efklides, 2008; Norman et al., 2019). This definition is in keeping with the work of Nelson and Naren (1990, 1994) and their early work in defining metacognition as the product of control and monitoring processes. Moreover, the type of monitoring or reflective thought under consideration is generally considered conscious and deliberative,[1] making it accessible to the individual as they engage mental activity (e.g., Koriat, 2007 and see Evans & Stanovich, 2013 and Lyons & Zelazo, 2011 for discussion of cognitive heuristics associated with Type II deliberative and reflective processes).

There is a strong empirical basis tied to noting age differences in the various skills that comprise cognitive monitoring. This body of work has largely entailed comparing cognitive monitoring skills in children at different ages (Bryce & Whitehead, 2012; Schneider et al., 2000; Schneider & Lockl, 2002) or the same child at different time points (e.g., Schneider et al., 2004). Experimentally speaking, the basic framework is to provide the subjects with various cognitive tasks asking them about what strategies they employ when thinking how to solve problems. Children of different ages are asked how they remember when they read a story, whether they repeat things in order to memorize, how they organize facts (i.e., mnemonics), and other mental organizational skills and actions that help them attend, memorize, and recall (e.g., Cross & Paris, 1988; Gascoine et al., 2017; Kurtz & Borkowski, 1984 and for a review of assessment strategies see, Baker & Cerro, 2000). This has spawned a" theory of mind" concept that can be used to differentiate cognitive monitoring at different ages (Flavell, 1988; Kuhn, 2000b; Wellman, 2018) and also when certain mental skills become more transparent to the individual (e.g., Pintrich, 2002). The emphasis on establishing development when cognitive monitoring occurs in the repertoire of youth is important. Cognitive monitoring can be used as a "thinking tool" to enable youth to plan and evaluate, giving them better problem-solving skills with real-world applications. This ability falls

[1]Flavell (1979) notes that metacognition can be unintentional and automatic prompted by retrieval or task-related cues. Although not the focus of this paper, we acknowledge that System I (intuitive thinking) can be responsible for eliciting automatic responses (i.e., mental heuristics) that instigate cognitive monitoring. It is worth noting that Flavell also felt that metacognitive experiences are more likely to occur in situations (tasks) that require conscious thought as one applies some type of planning, evaluation, and decision-making that leads to an internal form of quality control. Veenman et al. (2006) provide arguments both for and against that cognitive monitoring has to be conscious as opposed to more automatic and beneath the radar of consciousness (see Efklides, 2008; Fox & Riconscente, 2008; Rosenthal, 2000 for different perspectives on the role of conscious thought in metacognition).

under the broad catchall of "21st century skills" (e.g., Binkley et al., 2012; Greiff et al., 2014; Partnership for 21st Century Skills, 2007; Trilling & Fadel, 2009) and is part of Bloom's taxonomy of learning, which emphasizes problem-solving, along with critical thinking, creativity, and metacognition (Bloom et al., 1956).

Given this brief review, three issues come to mind when cognitive monitoring is discussed. First, what is the complexion of cognitive monitoring, in other words, what are the core competencies and skills that provide a basis for this type of strategic mental thinking? Along these lines, the individual skills that comprise cognitive monitoring overlap empirically and conceptually or can they be distinguished psychometrically. Second, is cognitive monitoring stage-like or it unfolds in a more progressive if not fluid manner. At heart here is whether the cognitive skills of a 14-year-old are uniquely different from those of a 15- or 16-year-old or that an older youth just applies more of the same skill more effectively if not more frequently (Kuhn, 2000a). Addressing this question can help address a broader issue of whether a Piagetan approach based on assimilation and accommodation is appropriate to understand the development of cognitive monitoring. Moreover, addressing this issue would help substantiate Flavell's perspective that metacognitive experiences, knowledge, and strategies accrete over time because of life experiences and increasing cognitive resources. Third, are there gender differences in cognitive monitoring? Boys and girls generally perform equally well in terms of scholastic achievement tests (Duckworth & Seligman, 2006; Matthews et al., 2009; Voyer & Voyer, 2014), report being equally engaged in school (Wang et al., 2011), and female students show a slight advantage with regard to scholastic grades (Voyer & Voyer, 2014). Conceivably, there may be underlying differences in cognitive monitoring that reflect gender-specific cognitive strategies. However, with few exceptions, very few studies have pinpointed with any precision whether any of these observed gender differences in metacognition or cognitive heuristics carry over to cognitive monitoring skills.

Focus of the current study

In the current study, we address these three issues using a single overarching analytic framework. The framework involves using multiple group latent-variable confirmatory factor analysis (CFA). The latent-variable component provides a means to establish the psychometric properties of cognitive monitoring as a hypothetical unobserved construct. Psychometrically speaking, this assumes that cognitive monitoring can be "quantified" and that we can infer cognitive monitoring from the joint probability distribution of observed (measurable) lower-order cognitive activities (e.g.,

Borsboom et al., 2003; Veenman et al., 2006). To our knowledge, very few studies have examined the invariance of the self-regulatory qualities of cognitive monitoring using confirmatory approaches and with younger age populations (for exceptions see, Gomez et al., 2021; Li et al., 2023). As we explain below, we hypothesize three facets of cognitive monitoring based on unique sets of cognitive activities. This approach enables us to test whether the observed measures that make up each construct differ in the strength of their contribution to their respective latent constructs by age and gender. In contrast to many studies of metacognition conducted in the past, in the current study we don't examine strategies specific to mathematics, language or reading, which capture cognitive monitoring in strictly scholastic domains. Instead, we focus on cognitive monitoring skills that can be interpreted as "control or self-regulatory processes" encompassing strategies related to decision-making (e.g., gathering information), internal self-talk (e.g., self-reward), and affective self-management (emotional control). These three aspects of cognitive monitoring are assumed to operate at a higher or more general level and can be applied to multiple domains of cognitive functioning both academic (e.g., solving math problems or reading comprehension) and otherwise (e.g., losing one's direction on a hiking trail). The correlation between these different skills is an indication of whether these cognitive activities are divergent or unified and whether they share some common ground outside of the specific strategy they encompass.

The multiple group component provides a means to assess the fit of an a-priori model across age and gender groups. These models use increasingly restrictive constraints, each one addressing a different and important facet of whether groups differ in cognitive monitoring, including its measurement properties, the way that students respond to the questions used to model cognitive monitoring, the relations among the different facets of cognitive monitoring, and between-group differences in the mean levels (intercepts) of the strategies hypothesized to capture cognitive monitoring. Because the items used to assess cognitive monitoring are imperfect, we also test whether the error variances are equivalent across subgroups. This is a benefit of using confirmatory procedures, which enables estimating the error variances as an independent component of the measurement model. As part of the model testing sequence, we also assess the fit of a higher-order structure that posits a single overarching latent variable with lower-order primary factors capturing unique (albeit not distinct) facets of cognitive monitoring. This latter model comes closest to the extant literature on "metacognition" and assumes that the lower-order cognitive processes reflect a single agent or mental force that selects from a pool of regulatory strategies fitting the cognitive activity.

Method

Data for this study was gathered as part of a 3-year longitudinal study of sports motivation and cognitive functioning in youth attending high schools in the Czech Republic. The Czech Republic is divided into 14 regions (corresponding to provinces) that were geographically merged with adjacent regions collapsed into five contiguous regions. Schools were randomly selected from these five areas with one large school (>450 students) and 1 to 2 small schools (<450 students) drawn from each region. Schools were approached individually through administrative contacts and asked if they wished to participate in longitudinal study. A total of six schools refused to participate and selection stopped after 13 schools (distributed relatively evenly throughout the five areas) agreed to participate. Consenting schools then announced the study (posting flyers and through homeroom announcements) and provided students with a hyperlink to the survey.

The initial welcome screen presented the main purpose of the research and students (if under 18 years of age, also their legal representatives) were then provided an assent form (in Czech language) they must read and electronically check their willingness to participate (Yes, I agree or No, I don't want to participate). The assent informs students (if under 18 years of age, also their legal representatives) of their rights as research participants, the voluntary nature of their participation, and that they can withdraw from the study at any point without penalty (IRB#142/22, Ethics Committee, Faculty of Physical Education and Sport, Charles University). The data collection period spanned from October 24th, 2024, to December 16th, 2024, allowing participants to complete the online surveys at any time during this interval.

Participation levels varied across schools and ranged from a low of 10% to a high of 86.7% (mean = 53.7%). The data that support the findings of this study are available from the corresponding author, upon reasonable request.

The online questionnaire included a set of demographic questions, questions about students' sports participation, quality of physical education facilities, school and teacher bonding, locus of control for physical education, peer and family support for sporting activities, depressive symptoms, optimism and pessimism, and the cognitive reflective questions that are the focus of the present study. Instruments that were not available in the Czech language, were translated by three independent individuals using acceptable translation and back translation procedures (e.g., Harkness et al., 2003; Walde & Völlm, 2023).

Measures

A total of 15 items were used to assess different facets of cognitive monitoring. Five items assessed decision-making and were taken from the Response Profile of the Coping Assessment Battery (Bugen & Hawkins, 1981; Wills, 1986). The items represent different theoretical perspectives with regard to applied and behavior-based coping including strategies individuals engage to obtain information, consider alternatives, weigh consequences, and evaluate one's options if a specific course of action is chosen (Pearlin & Schooler, 1978). These items are considered part of the vigilant coping style used in the 7-steps of effective decision-making (Janis & Mann, 1977). Internal consistency estimates for the five-item scale have been high in several longitudinal studies of adolescent samples (α's = .88–.91: Griffin et al., 2009; Scheier & Botvin, 1995; Scheier et al., 1997). Following a common stem ("When I have a problem, I …") students were provided five decision-making steps with sample items including "Think of as many possible choices or ways of solving the problem as I can" and "Think about what will happen for each choice before doing anything." Many of these items are quite similar to the monitoring and planning items contained in the 52-item Metacognitive Awareness Inventory (Schraw & Dennison, 1994).

Five items taken from the 30-item Self Reinforcement Scale (SRS; Heiby, 1982) assessed inner self-talk and reward strategies as a generalized response set for self-praise. Heiby developed the SRS as part of research examining rumination (covert) control of behavior through self-controlled reinforcement and its relations to depression (see for example, Lewinsohn, 1974 for more on the low frequency of self-reinforcement hypothesis). Sample items include ("When I do something well, I take time to enjoy the feeling" and "I silently praise myself for even small achievements"). Test-retest reliability (r = .92) and Spearman-Brown split-half reliability (r = .87) were both excellent for the full version when tested with young adults (Heiby, 1983). Several large-scale studies of adolescent development have reported excellent estimates of internal consistency calculated using McDonald's (1999) Omega (ω = .87, .88 for 8th and 10th grade youth: Griffin et al., 2021) and exceeding .85 when calculated using Cronbach's α (Griffin et al., 2002; Scheier & Botvin, 1995).

Five items were taken from the 36-item Self Control Schedule (SCS; Rosenbaum, 1980) to assess affective self-management. According to Rosenbaum, these are self-control strategies to manage anxiety and distress when confronting non-task specific situations. They encompass strategies that individuals use to minimize the effects of anxiety that arises from internal events and that can affect performance (e.g., take a deep breath and relax). The creation of the SCS was based on stress and coping

literatures with direct application to behavior therapies popular at the time (e.g., Lazarus & Folkman, 1984). The instrument contains multiple sections with one including 12 self-statements to control emotional and physiological intrusions. Sample items used in the current study include "If an unpleasant thought is bothering me, I try to think about something pleasant" and "When I am worried about something, I try to keep myself busy or think about other things." Omega for the five items that were retained based on adolescent samples was .90 (Griffin et al., 2021). Extensive factor analytic work accompanied all of the adolescent development studies resulting in briefer versions of the full-length scales accommodating their use in school-based surveys. Response scales for all 15 cognitive monitoring items ranged from 1 (*strongly disagree*) to 5 (*strongly agree*). S1 Table contains all 15 of the cognitive activity items and the abbreviations used throughout the article.

Analytic strategy

Model testing proceeded with a series of integrated steps that follows the sequential constraint imposition procedure recommended by Dimitrov (2010). First, we tested a multigroup confirmatory factor analysis (CFA) measurement model positing three individual latent factors assessing decision-making, self-reinforcement, and affective self-management skills for each age and gender group. We used the partial disaggregation scheme recommended by Bagozzi and Heatherton (1994) with individual items used as indicators of three separate latent factors. Factor intercorrelations were freely estimated. This configural invariance model addresses whether the same general factor structure (and pattern of fixed and free loadings) is supported in the different age and gender groups (i.e., the dimensions making up cognitive monitoring are the same across groups given the specified pattern of factor loadings). Thereafter, each subsequent model imposed increasingly restrictive levels of measurement invariance (MI). The next test imposes weak factorial invariance by constraining factor loadings (metric invariance) to equivalence. This model assesses whether the items function the same and have the same meaning in the different age and gender groups. As Dimitrov (2010) points out, equivalent factor loadings between groups essentially means that a one-unit change for the attribute in question in one group is equivalent to a one-unit change in the same attribute in another group (i.e., the measurement properties that determine the factor composition are identical across groups).

This test is followed by a model assessing scalar (strong) invariance setting the item intercepts to equivalence. This model assesses whether the

means of the items are identical across groups (i.e., do the students use the response scale in the same fashion) and is a prelude to interpreting latent mean differences between groups (Meredith, 1993). Weak invariance is assumed in the scalar model, as this is required to advance in the model testing procedure (e.g., Byrne et al., 1989). If intercepts are not equivalent between groups, it suggests an item-by-group interaction (Robitzsch & Lüdtke, 2023) and indicates the subgroups do not respond to the items in a similar fashion. The strictest form of invariance is then tested by constraining the item error variances (item-specific residual variance) to equality.[2] This test is followed by tests of construct-level invariance including constraints on the factor variances and then covariances (factor loadings were kept invariant in this model) and the equivalence of factor means (also called structural invariance). The model configuration sets the mean of one factor in one group to zero and allows the factor mean to vary in all other groups (by age and gender). This makes the parameter in the unconstrained group equal to the "difference" in latent means (relative to the reference group: e.g., Aiken et al., 1994; Steenkamp & Baumgartner, 1998). The examination of latent mean differences is substantively important and essential to determine if the hypothetical construct capturing cognitive monitoring "matures" developmentally as Flavell suggested it would. As required, both metric and scalar invariance are maintained in this model (e.g., Raykov et al., 2012).

As a final step in the process, we specified a higher-order model positing a general factor of "cognitive monitoring" based on the associations between the three primary factors. This is more of a theory-driven test consistent with the literature suggesting that a higher-order latent factor of cognitive monitoring consists of multiple primary factors reflecting individual cognitive strategies that support a general monitoring effort. The strategy used for testing invariance in the primary factor model is utilized for testing invariance in the second-order model, beginning with configural invariance, followed by metric invariance imposed on the primary factors (e.g., Marsh & Hocevar, 1985). Then, the factor loadings on the second-order factor were constrained to equality between groups. A third model imposed scalar invariance on the first-order factor intercepts.

[2]Many authors have commented that the restriction on item uniquenesses is not essential for invariance testing unless test development is a central concern. Residual variance is considered non-factor determined variance reflecting item-specific variation (i.e., test-specific method variance, omitted variables, random measurement error or a function of the item not captured by the latent construct). Here, we merely apply this step to comply with accepted conventions testing factorial invariance (e.g., Dimitrov, 2010; Marsh et al., 2009). Lubke and Muthén (2005) pointed out that equality of residual variances is not essential for factor mean comparisons, which only requires scalar invariance as a prerequisite. This is because the error residual variance is independent of the factor (orthogonal terms) and invariance of item residuals has no effect on the factor means (see also Putnick & Bornstein, 2016 for additional commentary).

Both absolute and incremental goodness-of-fit indices were used jointly to evaluate whether a model positing invariance constraints with more parameters improves over a model absent these constraints (Hu & Bentler, 1999). The fit indices included the root mean square error of approximation (RMSEA; Browne & Cudeck, 1992; Steiger, 1990), the comparative fit index (CFI; Bentler, 1990), and the ratio of χ^2/df. The CFI (ranging from 0 to 1), in particular, is not sensitive to sample size and assesses the discrepancy between the sample data and the implied population model for structured means and variance/covariances. Values of the RMSEA $\leq$ .06, CFI $>$ .95, and the normed χ^2/df $\leq$ 5.0 meet the accepted threshold criteria for adequate model fit. Given the imposition of increasingly restrictive restraints in the invariance testing procedures, the models are nested (one has more parameter restrictions than another) and can be contrasted statistically ($\Delta\chi^2 = \Delta\chi^2_{\text{constr.}} - \chi^2_{\text{unconstr.}}$) to address whether there is a decrement in fit.[3] Invariance at any of the levels is plausible if the $\Delta\chi^2$ is not significant at the specified p-value ($<$ .05). The RMSEA is not sensitive to sample size and it can be used to detect nested model fit if the 90% confidence intervals from two models overlap, the models are not statistically different. In addition, following conventions proposed by Cheung and Rensvold (2002), we also used the change in CFI (ΔCFI) $\leq$ .01 as a critical benchmark in determining whether a model positing invariance is acceptable.[4] This is because the $\Delta\chi^2$ is sensitive to sample size and may lead to Type I error rates exceeding the conventional nominal level (i.e., over-rejection of invariance tests). All models were tested using the Mplus statistical software (V8.11: Muthén & Muthén, 1998–2017) using maximum likelihood estimation with analysis of the mean and covariance structures. We did not consider the use of ML estimation with robust standard errors necessary because there was no evidence of meaningful violations of normality assumptions. Across the 15 indicators, average absolute skewness was .233 (range = −.426–.227) and average kurtosis was .291 (range = −.576–.234), well within commonly accepted thresholds for approximate normality. Importantly, maximum likelihood estimation has been shown to be robust to modest departures from (multivariate)

[3]The notion of nested models hinges on one model being a specialized form of the other. In the case of parameter nesting, one model has more free parameters constrained to equality (i.e., restricted model), but it is considered competing with a more relaxed model. Under the null hypothesis of equivalent models −2*log likelihood is asymptotically distributed as a χ^2 variate, with the Δdegrees of freedom computed as the difference between the two nested models.

[4]Dimitrov (2010) points out that a larger ΔCFI of $>$ .01 provides evidence of measurement invariance because a larger value of the CFI indicates better fit (i.e., more variance and covariances accounted for by the implied model). Chen (2007) proposes alternative goodness-of-fit indices to gauge model fit in invariance tests pairing the ΔRMSEA and ΔSRMR with the traditional criteria of ΔCFI $\leq$ .01. Across the many arguments made for using different statistics to gauge model fit, the $\Delta\chi^2$ remains the most accepted because it has a known sampling distribution.

normality, particularly with continuous indicators and adequate sample sizes, yielding unbiased parameter estimates and accurate inference under such conditions (Enders, 2001).

Results

Sample description

S2 Table provides an overview of the sample characteristics by grade and gender. Student ages conformed to the European model of K-12 education, which starts schooling one year later than most other countries (9th graders are ~15 years of age). Mean age for the entire sample was 16.62 (SD = 1.182). More than half of the sample (52.5%) was female and 94.3% were of Czech nationality, 2.3% were Ukrainian and the rest were Slovakian, Vietnamese, Roma or chose the Other race/ethnicity category (all 1%). Almost two-thirds of the sample (63%) reported they lived in a two-parent household and on average had 3.84 members in their household (SD = 1.10). With few exceptions, the demographic breakdowns were consistent for each age and gender group.

Results of the CFA models testing invariance

Figure 1 shows the results of the fully disaggregated three-factor CFA measurement model that was tested for each age (Part A) and gender (Part B) group[5] (i.e., configural invariance). As depicted, the loadings are reasonably similar between both age and gender groups. Importantly, all of the factor correlations were below unity (although high in some cases) reinforcing that there was an element of discriminant validity between the different facets of cognitive monitoring (e.g., Steenkamp & Baumgartner, 1998). Although there is no agreement on the objective criteria for configural invariance, it is frequently assumed that model fit provides a reasonable barometer of fit. In all cases, the model fit was adequate for the different age groups, $\chi^2(348) = 1790.946$, CFI = .906, RMSEA = 0.075, 90% CI (.072–.079) and gender, χ^2 (174) = 1611.745, CFI = .907, RMSEA = 0.075, 95CI (.072–.079). Essentially, this shows that the hypothesized structure of

[5]Preliminary psychometric analyses provided a base of information regarding the reliability of the items used to form the latent factors. This information suggested that a fully disaggregated model would be the most appropriate way to test the latent variable factor structure. S3 Table shows the results of Mokken's H coefficient and McDonald's Omega, reinforcing the integrity of the scales, and a clear unidimensional factor structure for each latent construct. Given that data was collected on a school basis, we computed intraclass correlations coefficients (ICCs) for the different measures of cognitive monitoring (using observed multi-item composites). The ICC is a measure of clustering and can indicate how much variance in the observed measures (factor indicators) can be attributed to school-level as opposed to individual-level effects (Donner & Koval, 1982). The ICCs ranged from a low of .00 for two items assessing self-regulation to a high of .02 for a single item assessing decision making skills and averaged .006 across all 15 indicators. This wealth of supporting psychometric evidence encouraged us to move directly to the CFA models purporting simple structure.

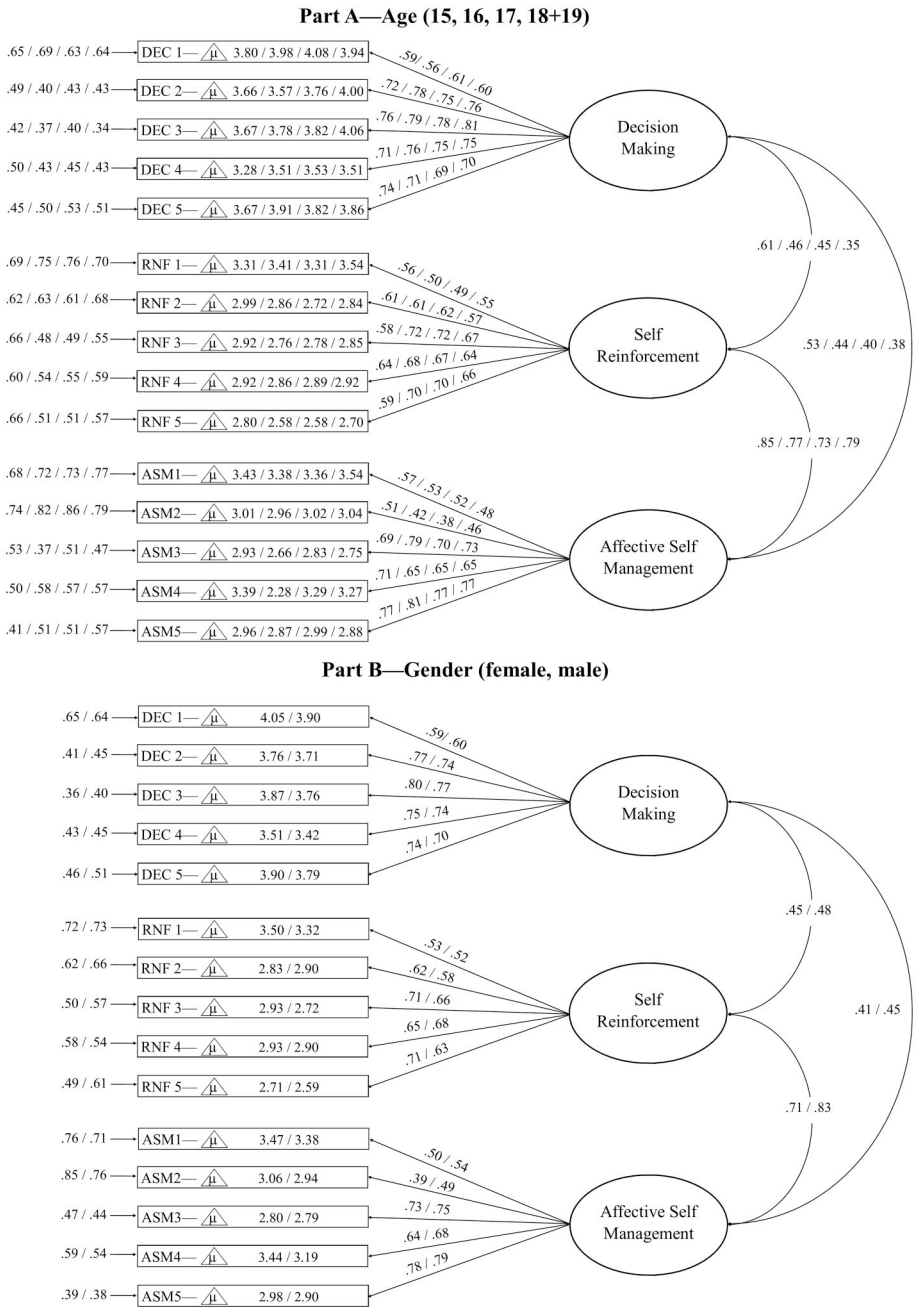

Figure 1. Standardized estimates for the three-factor CFA model for age (Part A) and gender (Part B).

cognitive monitoring (the pattern of fixed and free model parameters and 3-factor model configuration) can be reproduced by the sample data in the different subgroups.

Age differences

The first in a series of more restrictive models tested metric ("weak") invariance for age and separately gender groups. Turning first to the age models, the $\Delta\chi^2$ was nonsignificant indicating the model configuration positing equivalent loadings across the different age groups was tenable. The average factor loading (λ) for each latent construct and for each age group showed little variation .704, .719, .715, .725 for decision-making, .596, .642, .641, .616 for self-reinforcement, and .649, .639, .604, .619 for self-management, for 15, 16, 17, and 18–19 age groups, respectively. Moreover, as a type of "sensitivity analysis" the average absolute difference between all six possible combinations of age group factor loadings within construct was ${}_{\text{avg.}}\Delta\lambda = .019$, .016 and .026 for decision making, self-reinforcement, and affective self-management, respectively. The next model posits scalar ("strong") invariance for the intercepts of the different indicators. This model was compared to the metric invariance model and the nonsignificant $\Delta\chi^2$ again reinforced that MI of item intercepts across age groups was plausible. Table 1 also shows that the nested comparison of a model positing ("strict") invariance of error variances (capturing method variance or residual error) with the scalar model reinforced invariance. The next model in the MI sequence constrained the factor variances and also reinforced there was subgroup invariance.

The tests of invariance for the subgroupswere followed by a model at the construct level positing invariance of covariances. The nested comparison for this model (compared to the previous step) was significant, indicating there was noninvariance. Modification indices showed that freeing the covariance between decision-making and self-reinforcement for all four age groups and freely estimating the correlation between self-reinforcement and affective self-management for the 15- and 17-year-old age groups would improve the model fit.[6] Table 2 shows the partial invariance model for factor correlations fit well and the $\Delta\chi^2$ was nonsignificant when compared to the earlier model in the sequence specifying invariant factor variances. The association between decision-making and self-reinforcement was largest in magnitude for the youngest age group ($r = .538$) compared to the 16- ($r = .463$), 17- ($r = .463$), and 18+ year-old students ($r = .381$). The association between self-reinforcement and affective self-management was $r = .834$ for the 15-year-olds and $r = .728$ for the 17-year-olds. The average difference in the correlations within age group and between the fully invariant and partial invariant models was $\Delta r_{avg} = .006$ for decision-making and self-

[6]We used the sequential (backward) method of relaxing constraints based on the magnitude of the LaGrange modification indices (Yoon & Kim, 2014). This has a smaller Type I error rate as opposed to relaxing all of the constraints at once (the model is reconstituted with each relaxed constraint).

Table 1. Goodness-of-Fit Indices for Invariance Tests by Age and Gender Groups

Model	Against	$\chi^2(df)$	$\Delta\chi^2(\Delta df)^a$	p^a	CFI	ΔCFI^b	RMSEA (CI)
Age comparisons							
A. Configural invariance		1790.946 (348)			0.906	—	.075 (.072–.079)
B. Metric invariance	A	1836.952 (384)	46.01 (36)	0.123	0.905	0.001	.072 (.069–.075)
C. Scalar invariance	B	1879.375 (420)	42.42 (36)	0.214	0.905	0	.069 (.066–.072)
D. Error variances	C	1935.635 (465)	56.26 (45)	0.121	0.904	0.001	.066 (.063–.069)
E. Factor variances	D	1950.384 (474)	14.75 (9)	0.098	0.904	0	.065 (.062–.068)
F. Factor correlations invariance	E	1978.086 (483)	27.7 (9)	0.001	0.903	0.001	.065 (.062–.068)
F1. Partial factor correlation invariance	E	1959.397 (479)	9.01 (5)	0.109	0.904	0	.065 (.062–.068)
G. Factor means invariance	F1	1984.129 (488)	24.73 (9)	0.003	0.902	0.002	.065 (.062–.068)
G.1. Factor means partial invariance	F1	1968.991 (487)	9.59 (8)	0.295	0.903	0.001	.065 (.062–.068)
Gender comparisons							
A. Configural invariance		1611.745 (174)			0.907	—	.075 (.072–.079)
B. Metric invariance	A	1623.843 (186)	12.1 (12)	0.438	0.907	0	.073 (.070–.076)
C. Scalar invariance	B	1697.420 (198)	73.58 (12)	0.000	0.903	0.004	.072 (.069–.075)
C1. Partial scalar invariance	B	1629.137 (194)	5.29 (8)	0.726	0.907	0	.071 (.068–.075)
D. Error variances	C1	1683.716 (209)	54.58 (15)	0.000	0.904	0.003	.070 (.067–.073)
D1. Partial error invariance	C1	1639.459 (206)	10.32 (12)	0.588	0.907	0	.069 (.066–.072)
E. Factor variances	D1	1648.489 (209)	9.03 (3)	0.029	0.907	0	.069 (.066–.072)
E.1. Partial factor variance invariance	D1	1639.735 (208)	0.28 (2)	0.871	0.907	0	.069 (.066–.072)
F. Factor correlations invariance	E1	1662.228 (211)	22.49 (3)	0.000	0.906	0.001	.069 (.066–.072)
F1. Partial factor correlation invariance	E1	1642.634 (210)	2.9 (2)	0.235	0.907	0	.069 (.065–.072)
G. Factor means invariance	F1	1653.921 (213)	11.29 (3)	0.010	0.907	0	.068 (.065–.071)
G.1. Factor means partial invariance	F1	1645.235 (212)	2.6 (2)	0.272	0.907	0	.068 (.065–.071)

Note: non-significant *p*-value $>$ .05 indicates invariance across groups. Modification indices (MIs) above the critical value 3.84(1 *df*) were inspected for sources of invariance. CFI = comparative fit index, RMSEA = root mean square error of approximation; CI = 90% Confidence Interval. [a]$p < .05$ indicates significant difference in the nested model test and that the invariance constraint is not tenable. [b]ΔCFI $\leq$ .01 (indicates invariance). Model is base model and not nested. Sample sizes: Age: 15 = 619, 16 = 910, 17 = 717, 18 and older = 721. Gender: F = 1558, M = 1348. Fit for the full sample of the configural model: $\chi^2(87) = 1505.435$, $p < .001$, CFI = .910, RMSEA = .074 (CIs: .071–.077).

Table 2. Goodness-of-Fit Indices for Higher-Order Model Invariance Tests by Age and Gender Groups

Model	Against	$\chi^2(df)$	$\Delta\chi^2(\Delta df)^a$	p^a	CFI	ΔCFI[b]	RMSEA (CI)
Age comparisons							
A. Configural invariance		1935.635 (465)			0.904	—	.066 (.063–.069)
B. Metric invariance	A	1949.799 (471)	14.16 (6)	0.028	0.904	0	.066 (.063–.069)
B1. Partial metric invariance	A	1943.199 (470)	7.56 (5)	0.182	0.904	0	.066 (.063–.069)
C. Scalar invariance	B1	1960.38 (476)	17.18 (6)	0.009	0.903	0.001	.065 (.062–.068)
C1. Partial scalar invariance	B1	1946.975 (475)	3.78 (5)	0.582	0.904	0	.065 (.062–.068)
D. Factor variances	C1	1955.079 (478)	8.1 (3)	0.044	0.904	0	.065 (.062–.068)
D1. Partial factor variance invariance	C1	1950.277 (477)	3.3 (2)	0.192	0.904	0	.065 (.062–.068)
E. Factor means invariance	D1	1956.371 (480)	6.09 (3)	0.107	0.904	0	.065 (.062–.068)
Gender comparisons							
A. Configural invariance		1639.459 (206)	—	—	0.907	—	.069 (.066–.072)
B. Metric invariance	A	1643.937 (208)	4.48 (2)	0.107	0.907	0	.069 (.066–.072)
C. Scalar invariance	B	1654.232 (210)	10.3 (2)	0.006	0.906	0.001	.069 (.066–.072)
C1. Partial scalar invariance	B	1644.164 (209)	0.23 (1)	0.634	0.907	0	.069 (.066–.072)
E. Factor variances	C1	1650.607 (210)	6.44 (1)	0.011	0.907	0	.069 (.066–.072)
G. Factor means invariance	C1	1646.542 (210)	2.38 (1)	0.123	0.907	0	.069 (.066–.072)

Note: non-significant *p*-value $>$ .05 indicates invariance across groups. Modification indices (MIs) above the critical value 3.84(1 *df*) were inspected for sources of invariance. CFI = comparative fit index, RMSEA = root mean square error of approximation; CI = 90% Confidence Interval. [a]$p < .05$ indicates significant difference in the nested model test and that the invariance constraint is not tenable. [b]ΔCFI $>$ .01 indicates a lack of invariance. Sample sizes: Age: 15 = 619, 16 = 910, 17 = 717, 18 + 19 = 721. Gender: F = 1558, M = 1348.

reinforcement and $\Delta r_{avg} = .004$ for the correlation between self-reinforcement and affective self-management.

The next step in the MI procedures included testing invariance of factor means. Constraining factor means to equivalence also provided evidence of noninvariance and required that we relax one mean equivalence (decision-making) in the 15-year-old age group. This modification resulted in a non-significant *p*-value in the comparison of the partial invariance model to the partial invariant factor correlation model. The standardized mean difference for the 15-year-old age group was $\mu = -0.178$, which is significant by the *z*-critical test, $z = 3.872$ (SE = .025), $p < .001$. This can be interpreted as an effect size equivalent to Cohen's *d*. The negative sign indicates that the youngest group scored 0.178 standard deviation units lower than the reference group consisting of the oldest youth.[7]

Gender differences

The test of metric invariance produced a nonsignificant $\Delta\chi^2$ indicating equivalent factor loadings for males and female students ($_{avg.}\Delta\lambda = .02, .02,$ and .03 for the three primary factors, respectively). A model constraining intercepts (and factor loadings) was significantly different from the previous step in the model testing sequence, thus indicating some of the intercepts should be freely estimated in the two groups. Based on LaGrange modification indices, a model freeing four intercepts fit well (Table 1) and was not significantly different from the metric invariance model indicating that partial scalar invariance was obtained. In all four of the freely estimated intercepts, female studentsscored higher than male students. Following Chen (2008), the average absolute difference in intercepts between the fully constrained and partial invariance scalar model for girls was .02 and for boys was .001.

The next model in the sequence constrained error variances to equality across groups. This model incorporated any of the noninvariance elements from previous models (intercepts that were unequal were left to freely vary across groups). The model was significantly different from the partial scalar invariance model with the ΔCFI not meeting the critical threshold of .001. Modification indices indicated that three residual errors should be freed. The resulting model was compared to the partial scalar invariance model and provided evidence the freed parameters and remaining constraints positing equality of residual errors was plausible. Here too, the average

[7]Follow-up pairwise comparisons (*t*-tests) indicated that the 17- to 15-year-old comparison was significant, $\mu_{diff} = 0.115$ (SE = 0.033), *z*-critical value = 3.468, $p < .001$ and the comparison of 16- to 15-year-olds was significant, $\mu_{diff} = .095$ (SE = .031), z-critical = 3.006, $p < .01$, with both favoring older students scoring higher. These models were run with both the scalar invariance and the factor mean invariance configurations producing the same results.

absolute difference in error variances between the fully constrained and partial error invariance model for both female and male students was $\Delta\delta_{avg.} = .008$. These numbers reinforce that there was little disfigurement to the model estimates as a result of relaxing equality constraints. A model positing equality of factor variances indicated that one factor variance constraint (affective self-management) should be relaxed. The absolute average difference between factor variances within gender was $\Delta\sigma^2 = .009$ indicating little model distortion by freely estimating a single factor variance term.

The next model in the MI sequence was part of the construct-related invariance testing procedure and involved positing equality of factor covariances/correlations. This model indicated that one correlation between self-reinforcement and affective self-management needed to be freely estimated. The association was smaller in magnitude for females ($r = .721$) compared to male students (.823). The average absolute difference within gender for the three correlations comparing the full and partial invariance models was $\Delta r_{avg.} = .009$. The final step in the measurement invariance procedure for the three primary factors tested the equivalence of latent factor means. As Table 1 shows, this model required relaxing a single constraint on the mean of decision-making skills. The mean gender difference was significant by the z-critical test, $\mu_{diff} = .111$ (SE = .021), z-critical = 2.941, $p = .003$, the positive value indicating that female students scored higher than male students.

Higher-order model

The magnitude of correlations among the primary factors for the age and gender subgroups suggests that we may be able to specify a higher-order factor tapping cognitive monitoring representing one facet of metacognition. Figure 2, Parts A and B, shows the results of the higher-order model with a single factor hypothesized to statistically "cause" the associations among the first-order factors (factor loadings are shown separately for each age and gender subgroup).

The same sequence used in the primary factor invariance tests was repeated for the higher-order factor, testing a configural model, then constraining factor loadings and item intercepts across groups (constraints were imposed at the level of the primary factors). Table 2 shows the model fit indices corresponding to the higher-order MI tests for both age and gender.

Turning first to the age comparisons, the results of the configural model showed that the pattern of fixed and free loadings fit well in in the different age groups. The average loading for the three primary constructs was .806, .765, .737, and .728 for the 15, 16, 17, and 18-year-old age groups.

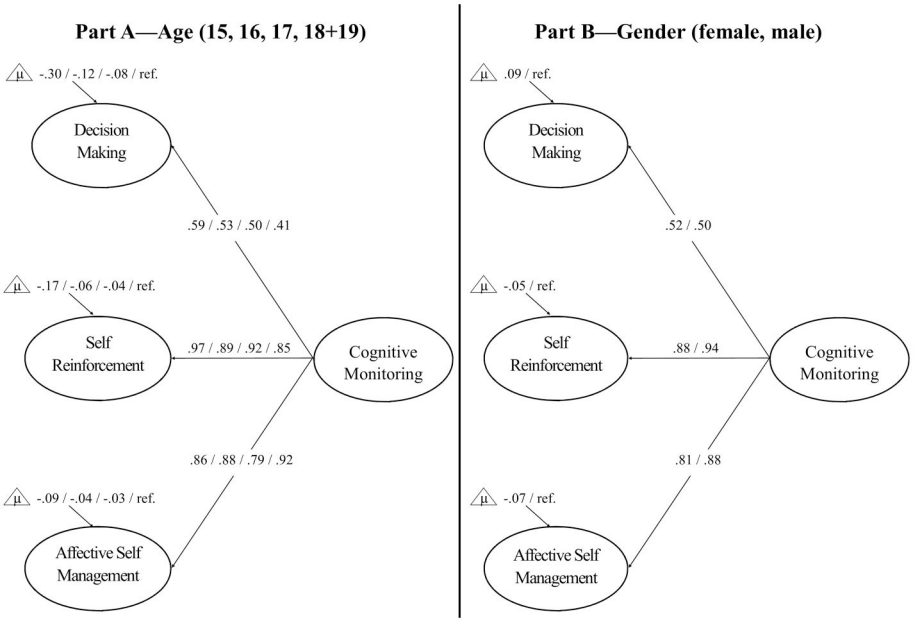

Figure 2. Standardized estimates for the higher order CFA model for age (Part A) and gender (Part B).

The average difference in factor loadings across all possible age comparisons was .09, .06, and .02 for decision-making, self-reinforcement, and affective self-management, respectively, suggesting very little variability in the magnitude of loadings from one age group to another. Referring to Figure 2A, the smallest magnitude loading was for decision-making skills at each age group (λ's = .587, .526, .501, and .413, respectively) and the largest magnitude of loading was for self-reinforcement (λ's = .974, .894, and .918 for the 15-, 16- and 17-year-old groups). The loading for affective self-management was largest for the 18–19 year-old group (λ = .924). The nested comparison between the metric and configural model indicated that a single factor loading (affective self-management) should be released for all of the age groups.[8] The resulting model showed the average difference in factor loading between the partial metric and full metric model was .003, .002, .033, and .008 within age group across the three primary factors.

A model positing invariant factor variance for the higher-order factor also showed evidence of noninvariance. The variance for the second-order construct was relaxed for the 16-year-old group and produced a nonsignificant $\Delta\chi^2$ (see Table 2). The final model in this sequence tested invariance of the higher-order factor mean across age groups. This model was not

[8]Since we obtained metric invariance in the primary factor models, we set the metric and identified the second-order model using the first primary factor (for both age and gender analyses).

significantly different from the previous step and thus no further testing occurred for the age subgroups.

The lower portion of Table 2 contains the measurement invariance testing sequence results for the gender higher-order model. The average loading for the primary factors was .737 for females and .774 for males ($\Delta\lambda$ = .04). The metric model provided evidence of invariance, but the scalar model was significantly different from the previous step indicating some of the primary factor means should be relaxed. Consistent with the primary factor model results, the intercept of decision-making was relaxed and this produced a good fitting model based on the $\Delta\chi^2$. The *z*-critical test indicated that females scored higher than boys, *z*-critical = 3.171 (SE = .038), $p < .01$, $\mu_{diff} = 0.122$. The factor variance model also showed the need to freely estimate the variance of the higher-order construct (σ^2 = .070 and .083 for females and male students, respectively). Factor means were invariant.

Discussion

In this study we took steps to clarify the composition and structure of the self-monitoring function of metacognition in four different age groups and by gender. Although there has been considerable discussion regarding the composition of cognitive monitoring and how it operates in different age and gender groups, very few studies have tested MI using quantifiable methods. This type of clarity is needed as cognitive monitoring is a central component of the 21st century skills that are being taught to youth across the world (e.g., Binkley et al., 2012; Greiff et al., 2014). These skills include competencies that are essential for critical thinking and problem solving, paving the way for learners to become leaders. The current findings show that at a very basic level, the self-regulatory function of cognitive monitoring is comprised of three uniquely different but related skills that collectively represent the procedural (tactical) elements of metacognition. These skills include how youth solve problems by gathering information, evaluating options and solutions, and weighing alternatives, the self-praise and encouragement strategies they use to reinforce themselves for small accomplishments, and the way they strategize to remain calm, address situations that may seem unpleasant, and quell anxiety before engaging a task. All represent forms of internal (self-regulatory) feedback students can engage as they monitor their progress in addressing tasks and encompass the behavioral, cognitive, and emotional tools that students implement to learn.

Tests of MI seek to establish whether the self-report items on a questionnaire mean the same thing to different groups (they interpret the questions the same), the different groups use the response scales in the same fashion,

test or item-specific error variances are consistent across groups, the pattern of factor covariances/variances are the same, and the factor means are the same (and dispersion is equivalent around the factor means). Beginning with the simplest comparison, the omnibus MI tests showed that the hypothesized multidimensional configuration of cognitive monitoring and the pattern of factor loadings is the same across age and gender groups. This means the three factors chosen to reflect cognitive monitoring (and the imposition of simple structure) were recovered from the variance-covariance matrices of the different subgroups. In addition, the imposition of metric invariance showed that the meaning of the constructs (i.e., magnitude of loadings) capturing the relationship of the measured variables to the latent constructs of cognitive monitoring (i.e., the regression weights) was equivalent for the different age and gender groups. This is an important step in the process of testing invariance because latent factors are not "real," rather they are hypothetical constructs inferred from the properties of the measured variables (e.g., Bollen, 2002). Thus, an important step in the invariance testing process relies on the veracity and validity of the first-order measurement model, which considers the pattern and magnitude of the factor loadings as they reflect the different facets of cognitive monitoring.

Tests of scalar invariance for the different age groups showed that students responded to the survey items assessing the different cognitive monitoring strategies in the same manner resulting in similar item means (i.e., equal scale intervals). Obtaining scalar invariance also means that the latent means are causing differences in the item intercepts without biasing them up or down in any manner (i.e., no additive bias or undue influence on the origin of the scales). A model imposing invariance of item means for male and female students provided evidence of item-to-group interactions reflecting differential item functioning. The four intercepts that were freely estimated indicated that female students had higher mean levels for vigilant and applied coping strategies.

Conceivably, one explanation for these differences suggests that female students benefit from socializing experiences where they are taught to mull things over, consider alternatives, and weigh the pros and cons of their decisions before acting. This type of strategic thinking would comport with findings that show females in this age group have better cognitive emotional regulation (e.g., Chaplin & Aidao, 2013; Sanchis-Sanchis et al., 2020), and they are better able to cope with stressful situations (e.g., Eschenbeck et al., 2007). Cognitive emotional regulation can include planning, refocusing, positive reappraisal, and putting things into perspective, some of which overlaps with the content of decision-making skills assessed in the current study.

There are also findings in the literature supporting more rumination in girls compared to boys in this age group (e.g., Dawson et al., 2023; Jose & Brown, 2008). Rumination (i.e., brooding, repetitive and persistent thoughts about failure) has been associated with depression even at young ages (e.g., Nolen-Hoeksema, 1994; Nolen-Hoeksema & Girgus, 1994). However, there is a brighter side to rumination because at times repetitive thinking may create more self-insight how to adapt one's cognitive strategies to the situation at hand. Rumination, in these terms, is best thought of as a form of coping (e.g., Broderick & Korteland, 2002) and may have contributed to the elevated scores in decision-making (a form of applied coping) in the female sample. Findings of this nature could point toward adolescent females being more "reserved," more cautious in their thinking (i.e., more pensive and less impulsive, see for example Shulman et al., 2015), as well as demonstrating better self-regulation (e.g., van Tetering et al., 2020) and coping strategies when faced with stress (e.g., Ziggert & Kistner, 2002). Carefully thinking about events, obtaining more information, weighing alternatives, and thinking about consequences are all formidable means to divert one's attention from negative events or experiences that is normally associated with the "dwelling or brooding" that are signature features of rumination.

Other explanations for the gender differences could be a response set reflecting social desirability, translation issues, or sample-specific behaviors that don't extend beyond the current sample of Czech high school students. In contrast to the item-level findings, we found no evidence of gender or age differences when imposing construct-level invariance including latent factor means. The latter finding suggests that the levels of cognitive monitoring, the interrelationships of the latent constructs, and the dispersion around the latent means were relatively similar across the different subgroups.

Overall, when nested model comparisons indicated noninvariance, only a relatively few parameters required relaxing. In any of the efforts to obtain partial invariance, we followed conventions suggested by Steenkamp and Baumgartner (1998), which entailed freely estimating parameters that had relatively large modification indices and also reparametrizing the model iteratively one parameter at a time to reconstitute the models. This procedure also avoids capitalizing on chance when making model modifications (MacCallum et al., 1992). Tests of strict invariance for age provided evidence of invariance, however, the same tests for strict invariance for gender required relaxing five constraints (33%) to achieve a better model fit. Apparently, there is no hard and fast rule (empirical or otherwise) on how many constraints can be released before there is a substantive change in the model (e.g., Byrne et al., 1989; Putnick & Bornstein, 2016). Monte

Carlo studies have indicated some bias on the parameter estimates with too many relaxed constraints (e.g., Steinmetz, 2013) although studies by Schmitt et al. (2011) and also Meade and Lautenschlager (2004) reported that models involving partial invariance contained little parameter estimate bias. We found very little bias by computing differences in the estimates before and after relaxing constraints, suggesting the models were not disturbed.

In both the case of age and gender we did not obtain invariance of covariances and variances, thus the correlations are not invariant across the subgroups. In the case of age, the association between decision-making and self-reinforcement was freely estimated in all of the age groups, and showed a pattern of decreasing in magnitude with increasing age. This means these skills are not as closely intertwined as youth gain rapport using them and they may show preference for using one set of skills over another. The association between self-reinforcement and affective self-management was freely estimated in the 15- and 17-year-old age groups. For the younger age group this association was larger and for the 17-year-old age group the association was smaller in magnitude compared to the older age groups. For gender, the correlation was smaller in magnitude for females between self-reinforcement and affective self-management, but the difference was quite trivial.

The rationale behind having factor variance/covariance invariance (and metric) is that it supports making cross-group comparisons using standardized measures. For instance, when correlating a factor that is invariant with an external marker factor, it is safer to make statements about discriminant validity when there is homogeneity in how the factor is conceptualized in different subgroups. Moreover, had we obtained weak, strong, and strict invariance, the items would be equally reliable (S3 Table shows that omega differed ever so slightly for the subgroups).

Even though we did not obtain complete invariance for all of the models tested, not too many parameters had to be freely estimated across the different subgroups. This bodes well for subgroup comparisons because all roads lead to being able to make quantitative comparisons of means across subgroups. In other words, by not having to tinker too much with the model configuration (i.e., relaxing too many constraints), we can revisit the issues raised by Flavell in his seminal article, and ask who is scoring highest on cognitive monitoring. In general, older students scored higher on the components of cognitive monitoring. More specific pairwise comparisons indicated that the 16- and 17-year-old age groups were significantly different (and higher) in the latent means of some of the facets of cognitive monitoring compared to the 15-year-old group and female students scored higher than male students on decision-making alone. Despite slight mean

differences in levels of cognitive monitoring, the results from all of the subgroup analyses point toward consistent measurement of cognitive monitoring as a psychometrically sound latent theoretical construct that can be inferred from the properties of measured variables.

The higher-order model produced additional information regarding the composition of cognitive monitoring and its consistency across age and gender groups. For instance, the configural model showed that the loading for decision-making skills was appreciably smaller in magnitude than the other two primary factors in all of the subgroups. This suggests that the meaning of the higher-order factor could be interpreted as reflecting more internal self-talk strategies targeting reward strategies and regulation of highly charged emotional states as opposed to behavioral coping skills that students use when faced with problem solving tasks. Notably, there were very small differences in the means of the primary factors and only a single relaxed factor variance (indicating some heterogeneity in the distribution of latent mean scores for 16-year-old group).

Limitations

There are a number of limitations to this study worth noting. The cross-sectional nature of the data prevents us from making causal inferences about cognitive monitoring that can be attributed to age and gender differences. Although we can address mean differences by age for between-group comparison, we cannot infer intra-individual change in the underlying cognitive skills. Understanding development in this manner requires prospective longitudinal data that follows the same individual over time, which we are in the process of gathering. The data represent cognitive monitoring in Czech high school students, and it may be worth pursuing cross-cultural comparisons. Many latent nuisance variables can contribute to the noted group differences including socialization, family factors, motivation, self-regulation, and other personality factors, all of which have been implicated in metacognition (e.g., Efklides, 2001; Harrison & Vallin, 2018). Additional studies implicate intelligence in the development of metacognition, but the literature is not conclusive on this matter based on both individual studies (Hertzog & Robinson, 2005; Veenman et al., 2004) and by a recent meta-analysis (Ohtani & Hisasaka, 2018). The literature on rumination suggest that girls think "inwardly" and ruminate about negative events more than boys, but we don't know if it is domain specific or extends to cognitive monitoring in terms of problem-solving and tasks conducted in more academic settings. As Veenman et al. (2006) point out, cognitive monitoring is complicated and not easy to study, both conceptually and methodologically. Numerous factors have to be considered, as students may actually possess

cognitive monitoring skills, but not use them for many reasons (e.g., test anxiety interferes with application). We did not examine the declarative component of metacognition (i.e., knowledge and experiences), only the procedural component, the latter owing to the focus on monitoring and self-regulatory skills (e.g., Schneider et al., 2022; Schraw, 1998).

Correlations between different cognitive skills may be inflated due to method variance, which can reflect a host of biases introduced by the participants' beliefs (implicit or otherwise) about cognitive monitoring (e.g., Podsakoff et al., 2003). The model testing process used large-sample statistical theory to determine model fit. It is likely that the models generated provided a plausible representation of the system of influences captured by the variables and their interrelationships. However, the ML goodness-of-fit indices are sensitive to sample size leading to an increased probability of rejecting a true model even with just trivial discrepancies in the residual matrix (Bentler & Bonett, 1980). Finally, we can rule out that any measurement differences are confounded by characteristics of the samples because we found so few significant differences in demographics for the different age and gender subgroups. Despite these concerns, the current study provides a more in-depth psychometric view of the self-regulatory aspect of cognitive monitoring and contributes to the growing literature by demonstrating whether its complexion varies along age and gender using a relatively large sample. Future studies may want to replicate this approach in an effort to establish the generality of findings.

Implications

Studies of cognitive monitoring essentially seek ways to improve students learning skills by teaching them ways to memorize, organize their thinking, and recall information using mnemonic strategies (e.g., Efklides, 2001; Pintrich, 2002). The goal of instructional efforts is to boost students' confidence they can access these strategies, and know when and how to apply them, in which situations, and for which specific tasks. Discerning these choices can render learning easier and also make it more enjoyable. There is now growing evidence through meta-analysis that programs of this nature work (e.g., Donker et al., 2014). Students are likely to become more self-aware of the conditions in which they should apply cognitive strategies, building greater self-knowledge, and achieve a sense of accomplishment. Several authors have commented that the ultimate goal of understanding cognitive monitoring is to glean more information about motivation and use this information to improve student scholastic performance (e.g., Efklides, 2008). In essence, understanding cognitive monitoring is essential to learning more about what prompts students to engage in tasks, remain persistent, and believe in their own abilities (Bandura, 1997). Once these

relationships are more clearly understood, efforts can be made to foster new cognitive monitoring skills and bolster students' judgments about their capabilities.

In the current study, and with relatively few differences, all of the hierarchical MI tests point toward both age and gender similarity in the measurement and structural composition of cognitive monitoring. This suggests that broad-based instructional programs that focus on improving cognitive monitoring skills can be applied across different age and gender subgroups with likely the same benefits overall. Still, there remain many questions about cognitive monitoring that need to be addressed. For instance, one of the important questions to address is whether improving cognitive monitoring strategies will improve student motivation in scholastic contexts and overall whether these improvements are linked to performance and self-knowledge. In other words, what is the link between internal self-reflection (i.e., introspection to become self-aware through cognitive monitoring) and motivation? Furthermore, what is the link between both cognition and motivation and self-knowledge in terms of traditional self-constructs including self-esteem and self-efficacy? Addressing these linkages can help to provide a basis for instructional methods that target improved cognitive monitoring skills as a means of addressing core issues of the self. In the current study, cognitive monitoring was comprised of three unique skills that tap into mental heuristics. It is of paramount importance then to learn whether improvements in one facet will rollover to other facets of monitoring. This type of skills-based cascade is conceivable because the learner or the "self" is at the core making decisions when to engage cognitive monitoring, when it is most likely to result in the desired endpoint.

There is also a great deal to be learned about when students fail to utilize their cognitive monitoring skills. This may be at the heart of why some students report they don't engage well in school or do poorly on tests, because they fail to implement cognitive monitoring skills that can provide impetus when faced with difficult tasks. The absence of "metaskills" may interfere with a student's willingness to engage in problem solving tasks (Mayer, 1998). To offset deficits in monitoring and regulation, programs emphasizing cognitive monitoring should focus on making students more attentive to their own self-knowledge (awareness of what strategies work and don't work), when to apply these strategies as heuristics to reduce cognitive "load," to be cognizant of the efficiency of strategies, and how to communicate what is happening in their interior world cogently. The ultimate goal is to make studying, learning, remembering, recalling, and problem solving all a bit more fun and easier.

Author notes

Martin Komarc, Ph.D., is an Assistant Professor at Charles University's Faculty of Physical Education and Sport and a Research Associate at its Institute of Biophysics and Informatics (First Faculty of Medicine), focusing on psychometrics and applied statistics.

Lawrence M. Scheier is a psychologist with interests in program development and evaluation. He is president of LARS Research Institute, a non-profit strategic planning and evaluation company and Senior Research Scientist at Prevention Strategies, Greensboro, NC. His research emphasizes the causes and consequences of drug use and evaluation of programs that promote positive youth adaptation.

Jana Novotná is a sport psychologist and PhD student specializing in athlete mental health, performance psychology, and well-being. She has experience working in elite sport and teaches at the university level. Her current research focuses on psychological factors related to performance, health, and quality of life.

Acknowledgements

We would like to thank all participating schools in this study for their help during the data gathering procedures.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by the Grantová Agentura České Republiky [Grant No. Project 23-05873S] and Charles University [Grant No. COOPERATIO SPOS]. The funding had no role in the study design, data collection, analysis, interpretation of data, writing of the manuscript, or the decision to submit it for publication.

Notes on contributors

Martin Komarc, Ph.D., is an Assistant Professor at Charles University's Faculty of Physical Education and Sport and a Research Associate at its Institute of Biophysics and Informatics (First Faculty of Medicine), focusing on psychometrics and applied statistics.

Lawrence M. Scheier is a psychologist with interests in program development and evaluation. He is president of LARS Research Institute, a non-profit strategic planning and evaluation company and Senior Research Scientist at Prevention Strategies, Greensboro, NC. His research emphasizes the causes and consequences of drug use and evaluation of programs that promote positive youth adaptation.

Jana Novotná is a sport psychologist and PhD student specializing in athlete mental health, performance psychology, and well-being. She has experience working in elite sport and teaches at the university level. Her current research focuses on psychological factors related to performance, health, and quality of life.

ORCID

Martin Komarc http://orcid.org/0000-0003-4106-5217
Lawrence M. Scheier http://orcid.org/0000-0003-2254-0123
Jana Novotná http://orcid.org/0009-0005-1202-4776

References

Aiken, L. S., Stein, J. A., & Bentler, P. M. (1994). Structural equation analyses of clinical subpopulation differences and comparative treatment outcomes: Characterizing the daily lives of drug addicts. *Journal of Consulting and Clinical Psychology*, *62*(3), 488–499. https://doi.org/10.1037/0022-006X.62.3.488

Bagozzi, R. P., & Heatherton, T. F. (1994). A general approach to representing multifaceted personality constructs: Application to state self-esteem. *Structural Equation Modeling: A Multidisciplinary Journal*, *1*(1), 35–67. https://doi.org/10.1080/10705519409539961

Baker, L., & Cerro, L. (2000). Assessing metacognition in children and adults. In G. Schraw & J. C. Impara (Eds.), *Issues in the measurement of metacognition* (pp. 99–145). Buros Institute of Mental Measurement.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. W. W. Freeman & Co.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*(2), 238–246. https://doi.org/10.1037/0033-2909.107.2.238

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin*, *88*(3), 588–606. https://doi.org/10.1037/0033-2909.88.3.588

Binkley, M., Erstad, O., Herman, J., et al. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Springer.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. In *Handbook I. Cognitive domain*. David McKay.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology*, *53*(1), 605–634. https://doi.org/10.1146/annurev.psych.53.100901.135239

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Brinthaupt, T. M. (2019). Individual differences in self-talk frequency: Social isolation and cognitive disruption. *Frontiers in Psychology*, *10*, 1088. https://doi.org/10.3389/fpsyg.2019.01088

Broderick, P. C., & Korteland, C. (2002). Coping style and depression in early adolescence: Relationships to gender, gender role, and implicit beliefs. *Sex Roles*, *46*(7–8), 201–213. https://doi.org/10.1023/A:1019946714220

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, *21*(2), 230–258. https://doi.org/10.1177/0049124192021002005

Bryce, D., & Whitebread, D. (2012). The development of metacognitive skills: Evidence from observational analysis of young children's behavior during problem-solving. *Metacognition and Learning*, *7*(3), 197–217. https://doi.org/10.1007/s11409-012-9091-2

Bugen, L. A., & Hawkins, R. C. (1981). *The Coping Assessment Battery: Theoretical and empirical foundations*. Paper presented at the meeting of the American Psychological Association.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*(3), 456–466. https://doi.org/10.1037/0033-2909.105.3.456

Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotional expression in children: A meta-analytic review. *Psychological Bulletin*, *139*(4), 735–765. https://doi.org/10.1037/a0030737

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology*, *95*(5), 1005–1018. https://doi.org/10.1037/a0013193

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cheung, G. W., & Rensvold, R. B. (1999). Testing factorial invariance across groups: A reconceptualization and proposed new method. *Journal of Management*, *25*(1), 1–27. https://doi.org/10.1177/014920639902500101

Cross, D. R., & Paris, S. G. (1988). Developmental and instructional analyses of children's metacognition and reading comprehension. *Journal of Educational Psychology*, *80*(2), 131–142. https://doi.org/10.1037/0022-0663.80.2.131

Dawson, G. C., Adrian, M., Chu, P., McCauley, E., & Stoep, A. V. (2023). Associations between sex, rumination, and depressive symptoms in late adolescence: A four-year longitudinal investigation. *Journal of Clinical Child and Adolescent Psychology*, *52*(5), 675–685. https://doi.org/10.1080/15374416.2021.2019049

Dimitrov, D. M. (2010). Testing for factorial invariance in the context of construct validation. *Measurement and Evaluation in Counseling and Development*, *43*(2), 121–149. https://doi.org/10.1177/0748175610373459

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, *20*(4), 391–409. https://doi.org/10.1007/s10648-008-9083-6

Donker, A. S., de Boer, H., Kostons, D., Dignath-van Ewijk, C. C., & van der Werf, M. P. C. (2014). Effectiveness of learning strategy instruction on academic performance: A meta-analysis. *Educational Research Review*, *11*, 1–26. https://doi.org/10.1016/j.edurev.2013.11.002

Donner, A., & Koval, J. J. (1982). Design considerations in the estimation of intraclass correlation. *Annals of Human Genetics*, *46*(3), 271–277. https://doi.org/10.1111/j.1469-1809.1982.tb00718.x

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, *98*(1), 198–208. https://doi.org/10.1037/0022-0663.98.1.198

Efklides, A. (2011). Interactions of metacognition and motivation and affect in self-regulated learning: The MASRL model. *Educational Psychologist*, *46*(1), 6–25. https://doi.org/10.1080/00461520.2011.538645

Efklides, A. (2008). Metacognition: Defining its facets and levels of functioning in relation to self-regulation and co-regulation. *European Psychologist*, *13*(4), 277–287. https://doi.org/10.1027/1016-9040.13.4.277

Efklides, A. (2001). Metacognitive experiences in problem solving.: Metacognition, motivation, and self-regulation. In A. Efklides, J. Kuhl, & R. M. Sorrentino (Eds.), *Trends and prospects in motivation research* (pp. 297–323). Kluwer Academic Publishers.

Enders, C. K. (2001). A primer on maximum likelihood algorithms available for use with missing data. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(1), 128–141. https://doi.org/10.1207/S15328007SEM0801_7

Eschenbeck, H., Kohlmann, C.-W., & Lohaus, A. (2007). Gender differences in coping strategies in children and adolescents. *Journal of Individual Differences*, *28*(1), 18–26. https://doi.org/10.1027/1614-0001.28.1.18

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223–241. https://doi.org/10.1177/1745691612460685

Flavell, J. H. (1988). The development of children's knowledge about the mind: From cognitive connections to mental representations. In J. Astington, P. Harris, & D. Olson (Eds.), *Developing theories of mind* (pp. 244–267). Cambridge University Press.

Flavell, J. H. (1979). Metacognition and cognitive monitoring: A new area of cognitive-developmental inquiry. *American Psychologist*, *34*(10), 906–911. https://doi.org/10.1037/0003-066X.34.10.906

Fox, E., & Riconscente, M. (2008). Metacognition and self-regulation in James, Piaget, and Vygotsky. *Educational Psychology Review*, *20*(4), 373–389. https://doi.org/10.1007/s10648-008-9079-2

Gascoine, L., Higgins, S., & Wall, K. (2017). The assessment of metacognition in children aged 4-16 years: A systematic review. *Review of Education*, *5*(1), 3–57. https://doi.org/10.1002/rev3.3077

Gomez, C. M. A., de Aranjo, J., & Castillo-Diaz, M. A. (2021). Testing the invariance of the Metacognitive Monitoring Test. *Psico-USF Braganca Paulista*, *26*(4), 685–696. https://doi.org/10.1590/1413-82712021260407

Greiff, S., Wüstenberg, S., Csapó, B., Demetriou, A., Hautamäki, J., Graesser, A. C., & Martin, R. (2014). Domain-general problem solving skills and education in the 21st century. *Educational Research Review*, *13*, 74–83. https://doi.org/10.1016/j.edurev.2014.10.002

Griffin, K. W., Scheier, L. M., Botvin, G. J., & Komarc, M. (2021). Adolescent transitions in self-management skills and relations to young adult alcohol use. *Evaluation & The Health Professions*, *44*(1), 25–41. https://doi.org/10.1177/0163278720983432

Griffin, K. W., Botvin, G. J., Scheier, L. M., Epstein, J. A., & Doyle, M. M. (2002). Personal competence skills, distress, and well-being as determinants of substance use in a predominantly minority urban adolescent sample. *Prevention Science*, *3*(1), 23–33. https://doi.org/10.1023/A:1014667209130

Griffin, K. W., Scheier, L. M., & Botvin, G. J. (2009). Developmental trajectories of self-management skills and adolescent substance use. *Health and Addictions*, *9*(1), 15–37. https://doi.org/10.21134/haaj.v9i1.48

Harkness, J. A., Van de Vijver, F. J. R., & Mohler, P. (2003). *Cross-cultural survey methods*. Wiley.

Harris, K. R. (1990). Developing self-regulated learners: The role of private speech and self-instruction. *Educational Psychologist*, *25*(1), 35–49. https://doi.org/10.1207/s15326985ep2501_4

Harrison, G. M., & Vallin, L. M. (2018). Evaluating the Metacognitive Awareness Inventory using empirical factor-structure evidence. *Metacognition and Learning*, *13*(1), 15–38. https://doi.org/10.1007/s11409-017-9176-z

Heiby, E. M. (1982). A self-reinforcement questionnaire. *Behaviour Research and Therapy*, *20*(4), 397–401. https://doi.org/10.1016/0005-7967(82)90100-0

Heiby, E. M. (1983). Assessment of frequency of self-reinforcement. *Journal of Personality and Social Psychology*, *44*(6), 1304–1307. https://doi.org/10.1037/0022-3514.44.6.1304

Hertzog, C., & Robinson, A. E. (2005). Metacognition and Intelligence. In O. Wilhelm & R.W. Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 101–121). Sage.

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

Janis, I. L., & Mann, L. (1977). *Decision making: A psychological analysis of conflict choice, and commitment*. Free Press.

Jose, P. E., & Brown, I. (2008). When does the gender difference in rumination begin? Gender and age differences in the use of rumination by adolescents. *Journal of Youth and Adolescence*, *37*(2), 180–192. https://doi.org/10.1007/s10964-006-9166-y

Keating, D. P. (1990). Adolescent thinking. In S. Feldman & G. Elliott (Eds.), *At the threshold: The developing adolescent* (pp. 54–89). Harvard University Press.

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction*, *22*(4), 296–298. https://doi.org/10.1016/j.learninstruc.2012.01.002

Koriat, A. (2007). Metacognition and consciousness. In P. D. Zelazo, M. Moscovitch, & E. Thompson (Eds.), *The Cambridge handbook of consciousness* (pp. 289–325). Cambridge University Press.

Ku, K. Y. L., & Ho, I. R. (2010). Metacognitive strategies that enhance critical thinking. *Metacognition and Learning*, *5*(3), 251–267. https://doi.org/10.1007/s11409-010-9060-6

Kuhn, D. (2000a). Metacognitive development. *Current Directions in Psychological Science*, *9*(5), 178–181. https://doi.org/10.1111/1467-8721.00088

Kuhn, D. (2000b). Theory of mind, metacognition, and reasoning: A life-span perspective. In P. Mitchell & K. J. Riggs (Eds.), *Children's reasoning and the mind* (pp. 301–326). Psychology Press.

Kurtz, B. E., & Borkowski, J. G. (1984). Children's metacognitions: Exploring relations among knowledge, process, and motivational variables. *Journal of Experimental Child Psychology*, *37*(2), 335–354. https://doi.org/10.1016/0022-0965(84)90008-0

Lazarus, R. S., & Folkman, S. (1984). *Stress, appraisal, and coping*. Springer.

Lewinsohn, P. (1974). A behavioral approach to depression. In R. Friedman & M, Katz (Eds.), *The psychology of depression: Contemporary theory and research* (pp. 157–185). John Wiley & Sons.

Li, F., Yuan, D., Gao, C., Xiong, K., Geng, F., & Zhang, L. (2023). Validity and reliability of the Metacognitions Questionnaire-30 (MCQ030) among Chinese adolescents. *Child Psychiatry & Human Development*, *56*(4), 1031–1040. https://doi.org/10.1007/s10578-023-01625-7

Lubke, G. H., & Muthén, B. O. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods*, *10*(1), 21–39. https://doi.org/10.1037/1082-989X.10.1.21

Lyons, K. E., & Zelazo, P. D. (2011). Monitoring, metacognitions, and executive function: Elucidating the role of self-reflection in the development of self-regulation. *Advances in*

Child Development and Behavior, *40*, 379–412. https://doi.org/10.1016/B978-0-12-386491-8.00010-4

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490–504. https://doi.org/10.1037/0033-2909.111.3.490

Marsh, H. W., & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept. First- and higher-order factor models and their invariance across groups. *Psychological Bulletin*, *97*(3), 562–582. https://doi.org/10.1037/0033-2909.97.3.562

Marsh, H. W., Muthén, B. O., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 439–476. https://doi.org/10.1080/10705510903008220

Matthews, J. S., Ponitz, C. C., & Morrison, F. J. (2009). Early gender differences in self-regulation and academic achievement. *Journal of Educational Psychology*, *101*(3), 689–704. https://doi.org/10.1037/a0014240

Mayer, R. E. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science*, *26*(1-2), 49–63. https://doi.org/10.1023/A:1003088013286

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Meade, A. W., & Lautenschlager, G. J. (2004). A Monte-Carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *11*(1), 60–72. https://doi.org/10.1207/S15328007SEM1101_5

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*(4), 525–543. https://doi.org/10.1007/BF02294825

Muthén, L. K., & Muthén, B. O. (1998–2017). *Mplus user's guide* (8th ed.). Muthén & Muthén.

Nelson, T. O., & Naren, L. (1990). Metamemory: A theoretical framework and new findings. In G. Bower (Ed.), *The psychology of learning and motivation* (Vol. 26, pp. 125–173). Academic Press.

Nelson, T. O., & Naren, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). MIT Press.

Nolen-Hoeksema, S. (1994). An interactive model of emergence of gender differences in depression in adolescence. *Journal of Research on Adolescence*, *4*(4), 519–534. https://doi.org/10.1207/s15327795jra0404_5

Nolen-Hoeksema, S., & Girgus, J. S. (1994). The emergence of gender differences in depression during adolescence. *Psychological Bulletin*, *115*(3), 424–443. https://doi.org/10.1037/0033-2909.115.3.424

Norman, E., Pfuhl, G., Sæle, R. G., Svartdal, F., Låg, T., & Dahl, T. I. (2019). Metacognition in psychology. *Review of General Psychology*, *23*(4), 403–424. https://doi.org/10.1177/1089268019883821

Ohtani, K., & Hisasaka, T. (2018). Beyond intelligence: A meta-analytic review of the relationship among metacognition, intelligence, and academic performance. *Metacognition and Learning*, *13*(2), 179–212. https://doi.org/10.1007/s11409-018-9183-8

Partnership for 21st Century Skills. (2007). *Framework for 21st-century learning*. http://www.p21.org/documents/P21_Framework_Definitions.pdf

Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciu, M., & Lœvenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relations to self-monitoring. *Behavioural Brain Research*, *261*, 220–239. https://doi.org/10.1016/j.bbr.2013.12.034

Pintrich, P. R. (2002). The role of metacognitive knowledge in learning, teaching, and assessing. *Theory Into Practice*, *41*(4), 219–225. https://doi.org/10.1207/s15430421tip4104_3

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review: DR*, *41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Pearlin, L. I., & Schooler, C. (1978). The structure of coping. *Journal of Health and Social Behavior*, *19*(1), 2–21. https://doi.org/10.2307/2136319

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *The Journal of Applied Psychology*, *88*(5), 879–903. https://doi.org/10.1037/0021-9010.88.5.879

Raykov, T., Marcoulides, G. A., & Li, C.-H. (2012). Measurement invariance for latent constructs in multiple populations: A critical view and refocus. *Educational and Psychological Measurement*, *72*(6), 954–974. https://doi.org/10.1177/0013164412441607

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 859–870. https://doi.org/10.1080/10705511.2023.2191292

Roebers, C. M. (2017). Executive function and metacognition: Towards a unifying framework of cognitive self-regulation. *Developmental Review*, *45*, 31–51. https://doi.org/10.1016/j.dr.2017.04.001

Rosenbaum, M. (1980). A schedule for assessing self-control behaviors: Preliminary findings. *Behavior Therapy*, *11*(1), 109–121. https://doi.org/10.1016/S0005-7894(80)80040-2

Rosenthal, D. M. (2000). Consciousness, content, and metacognitive judgments. *Consciousness and Cognition*, *9*(2 Pt 1), 203–214. https://doi.org/10.1006/ccog.2000.0437

Sanchis-Sanchis, A., Grau, M. D., Moliner, A.-R., & Morales-Murillo, C. P. (2020). Effects of age and gender in emotional regulation of children and adolescents. *Frontiers in Psychology*, *11*, 946. https://doi.org/10.3389/fpsyg.2020.00946

Scheier, L. M., & Botvin, G. J. (1995). Effects of early adolescent drug use on cognitive efficacy in early-late adolescence: A developmental structural model. *Journal of Substance Abuse*, *7*(4), 379–404. https://doi.org/10.1016/0899-3289(95)90011-X

Scheier, L. M., Botvin, G. J., & Baker, E. (1997). Risk and protective factors as predictors of adolescent alcohol involvement and transitions in alcohol use: A prospective analysis. *Journal of Studies on Alcohol*, *58*(6), 652–667. https://doi.org/10.15288/jsa.1997.58.652

Schmitt, N., Golubovich, J., & Leong, F. T. L. (2011). Impact of measurement invariance on construct correlations, mean differences and relations with external correlates: An illustrative example using Big Five and RIASEC measures. *Assessment*, *18*(4), 412–427. https://doi.org/10.1177/1073191110373223

Schneider, W., Kron, V., Hünnerkopf, M., & Krajewski, K. (2004). The development of young children's memory strategies: Findings from the Würzburg longitudinal memory study. *Journal of Experimental Child Psychology*, *88*(2), 193–209. https://doi.org/10.1016/j.jecp.2004.02.004

Schneider, W., & Lockl, K. (2002). The development of metacognitive knowledge in children and adolescents. In T. J. Perfect, & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 224–257). Cambridge University Press.

Schneider, W., Tibken, C., & Richter, T. (2022). The development of metacognitive knowledge from childhood to young adulthood: Major trends and educational implications.

Advances in Child Development and Behavior, *63*, 273–307. https://doi.org/10.1016/bs.acdb.2022.04.006

Schneider, W., Visé, M., Lockl, K., & Nelson, T. O. (2000). Developmental trends in children's memory monitoring: Evidence from a judgment-of-learning task. *Cognitive Development*, *15*, 115–134. https://doi.org/10.1016/S0885-2014(00)00024-1

Schraw, G. (1998). Promoting general metacognitive awareness. *Instructional Science*, *26*(1–2), 113–125. https://doi.org/10.1023/A:1003044231033

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, *19*(4), 460–475. https://doi.org/10.1006/ceps.1994.1033

Schraw, G., & Gutierrez, A. P. (2015). Metacognitive strategy instruction that highlights the role of monitoring and control processes. In A. Pena-Ayala (Ed.), *Metacognition: Fundaments, applications, and trends: A profile of the current state-of-the-art* (pp. 3–16). Springer. https://doi.org/10.1007/978-3-319-11062-2_1

Shulman, E. P., Harden, K. P., Chein, J. M., & Steinberg, L. (2015). Sex differences in the developmental trajectories of impulse control and sensation seeking. *Journal of Youth and Adolescence*, *44*(1), 1–17. https://doi.org/10.1007/s10964-014-0116-9

Steenkamp, J.-B., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*(1), 78–107. https://doi.org/10.1086/209528

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*(2), 173–180. https://doi.org/10.1207/s15327906mbr2502_4

Steinmetz, H. (2013). Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology*, *9*(1), 1–12. https://doi.org/10.1027/1614-2241/a000049

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. Jossey Bass.

van Tetering, M. A. J., van der Laan, A. M., de Kogel, C. H., de Groot, R. H. M., & Jolles, J. (2020). Sex differences in self-regulation in early, middle and late adolescence: A large-scale cross-sectional study. *Plos One*, *15*(1), e0227607. https://doi.org/10.1371/journal.pone.0227607

Veenman, M. V. J., Kok, R., & Blöte, A. W. (2005). The relation between intellectual and metacognitive skills in early adolescence. *Instructional Science*, *33*(3), 193–211. https://doi.org/10.1007/s11251-004-2274-8

Veenman, M. V. J., Van Hout-Wolters, B. H. A. M., & Afflerbach, P. (2006). Metacognition and learning: Conceptual and methodological considerations. Metacognition. *Metacognition and Learning*, *1*(1), 3–14. https://doi.org/10.1007/s11409-006-6893-0

Veenman, M. V. J., Wilhelm, P., & Beishuizen, J. J. (2004). The relation between intellectual and metacognitive skills from a developmental perspective. *Learning and Instruction*, *14*(1), 89–109. https://doi.org/10.1016/j.learninstruc.2003.10.004

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Walde, Pl., & Völlm, B. A. (2023). The TRAPD approach as a method for questionnaire translation. *Frontiers in Psychiatry*, *14*, 1199989. https://doi.org/10.3389/psyt.2023.1199989

Wang, M.-T., Willett, J. B., & Eccles, J. S. (2011). The assessment of school engagement: Examining dimensionality and measurement invariance by gender and race/ethnicity. *Journal of School Psychology*, *49*(4), 465–480. https://doi.org/10.1016/j.jsp.2011.04.001

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, *15*(6), 728–755. https://doi.org/10.1080/17405629.2018.1435413

Wills, T. A. (1986). Stress and coping in early adolescence: Relationships to substance use in urban school samples. *Health Psychology*, *5*(6), 503–529. https://doi.org/10.1037/0278-6133.5.6.503

Yoon, M., & Kim, E. S. (2014). A comparison of sequential and nonsequential specification searches in testing factorial invariance. *Behavior Research Methods*, *46*(4), 1199–1206. https://doi.org/10.3758/s13428-013-0430-2

Ziggert, D. L., & Kistner, J. A. (2002). Response styles theory: Downward extension to children. *Journal of Clinical Child and Adolescent Psychology*, *31*, 325–334. https://doi.org/10.1207/S15374424JCCP3103_04