

KEY POINTS

- Sixty-five percent of multiple choice questions across four exams had at least one flaw that may introduce construct-irrelevant variance.
- Findings suggest item flaws should not be treated equal and that they most likely impact the item difficulty and ability level being tested. In addition, the type of flaw can have different effects for the examinee.

INTRODUCTION

- High-quality assessments are created by minimizing construct-irrelevant variance in multiple choice examinations.<sup>1</sup>
- Limited evidence exists to support guidelines that have identified item writing flaws (IWFs) that should be avoided, such as:<sup>2-4</sup>
  - using *all-of-the-above* or *none-of-the-above* answer choices,
  - including negative phrasing in items (e.g. *EXCEPT*, *NOT*), or
  - having heterogeneous content and grammatical structure.
- **Purpose:** Describe the influence of IWFs on the psychometric properties of multiple choice test items.

METHODS

IWF Checklist Design & Data Collection

- A **21-question IWF checklist** was developed, piloted, and refined by the research team based on multiple choice item writing guidelines.<sup>2</sup>
- **Seventy-two exam items** were compiled from 4 exams given to 1<sup>st</sup> year doctor of pharmacy students. Only standard multiple choice items were included (i.e., not true/false, K-type, etc.). Test questions were evaluated by 3 researchers using the checklist to identify IWFs.
- Blinded student performance data (n = 144) was obtained through course instructors. Item performance was dichotomously scored (i.e., incorrect response = 0 points, correct response = 1 point).

Data Analysis

- Calculated item difficulty, discrimination, internal consistency, and student performance based on **classical test theory (CTT)**.
- Compared CTT findings to the item discrimination and difficulty indices, item information, and student ability levels calculated using a 2-parameter (2P) **item response theory (IRT)** model.<sup>5</sup>
- Summarized psychometric properties by number and type of flaws.

RESULTS

TABLE 1 Item indices based on number of flaws [N=144; mean(SD)\* or median(IQR)]

	CTT Discrimination	IRT a Parameter	CTT Difficulty	IRT b Parameter	Max Information	Max Ability (Theta)
All Items (n=69)*	0.25 (0.13)	0.77 (0.56)	0.79 (0.17)	-2.32 (6.68)	0.23 (0.26)	-2.11 (2.20)
Zero Flaws (n=24)	0.25 (0.14–0.31)	0.61 (0.34–0.87)	0.76 (0.64–0.86)	-1.59 (-2.57– -0.92)	0.11 (0.04–0.19)	-1.60 (-2.57– -0.92)
One Flaw (n=18)	0.28 (0.20–0.34)	0.89 (0.68–1.29)	0.90 (0.72–0.94)	-2.41 (-3.03– -1.30)	0.20 (0.12–0.42)	-2.41 (-3.04– -1.31)
Two Flaws (n=15)	0.26 (0.20–0.36)	0.78 (0.44–1.01)	0.90 (0.76–0.91)	-2.81 (-3.53– -1.91)	0.15 (0.05–0.26)	-2.81 (-3.53– -1.91)

TABLE 2 Item indices based on type of flaw [N=144; median(IQR)]

	CTT Discrimination	IRT a Parameter	CTT Difficulty	IRT b Parameter	Max Information	Max Ability (Theta)
Flaw Helps Examinee (n=8)	0.26 (0.23–0.29)	0.82 (0.70–1.08)	0.91 (0.76–0.93)	-2.82 (-3.47– -2.41)	0.17 (0.12–0.30)	-2.82 (-3.47– -2.41)
Flaw Hurts Examinee (n=18)	0.26 (0.20–0.36)	0.75 (0.54–1.13)	0.87 (0.72–0.92)	-2.32 (-3.65– -1.20)	0.14 (0.08–0.32)	-2.32 (-3.65– -1.20)
Balance of Flaws (n=11)	0.26 (0.16–0.38)	0.90 (0.20–1.04)	0.90 (0.72–0.90)	-2.73 (-3.40– -1.82)	0.20 (0.01–0.27)	-2.73 (-3.40– -1.82)

CONCLUSION

- Preliminary evidence suggests the number and type of IWFs influence the quality of multiple choice questions and should not be considered equally problematic.
- IRT is a viable methodology that should be considered in addition to CTT when evaluating the impact of IWFs.
- Systematic research is needed to better quantify the impact of IWFs on the psychometric properties of multiple choice test questions.

REFERENCES


1. Downing SM. Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make a difference? *Acad Med* 2002;77(10):S103-S104.

2. Haladyna TM, Rodriguez MC. Developing and validating test items. New York, NY: Routledge; 2013.

3. Downing SM. The effects of violating standard item writing principles on tests and students: The consequences of using flawed items on achievement examination in medical education. *Adv in Health Sci Educ* 2005;10:133-143.

4. Pais J, et al. Do item-writing flaws reduce examinations psychometric quality? *BMC Res Notes* 2016;9:299-345.

5. Embretson SE, Reise SP. Item response theory for psychologists. Mahwah, NJ: Lawrence; 2000.



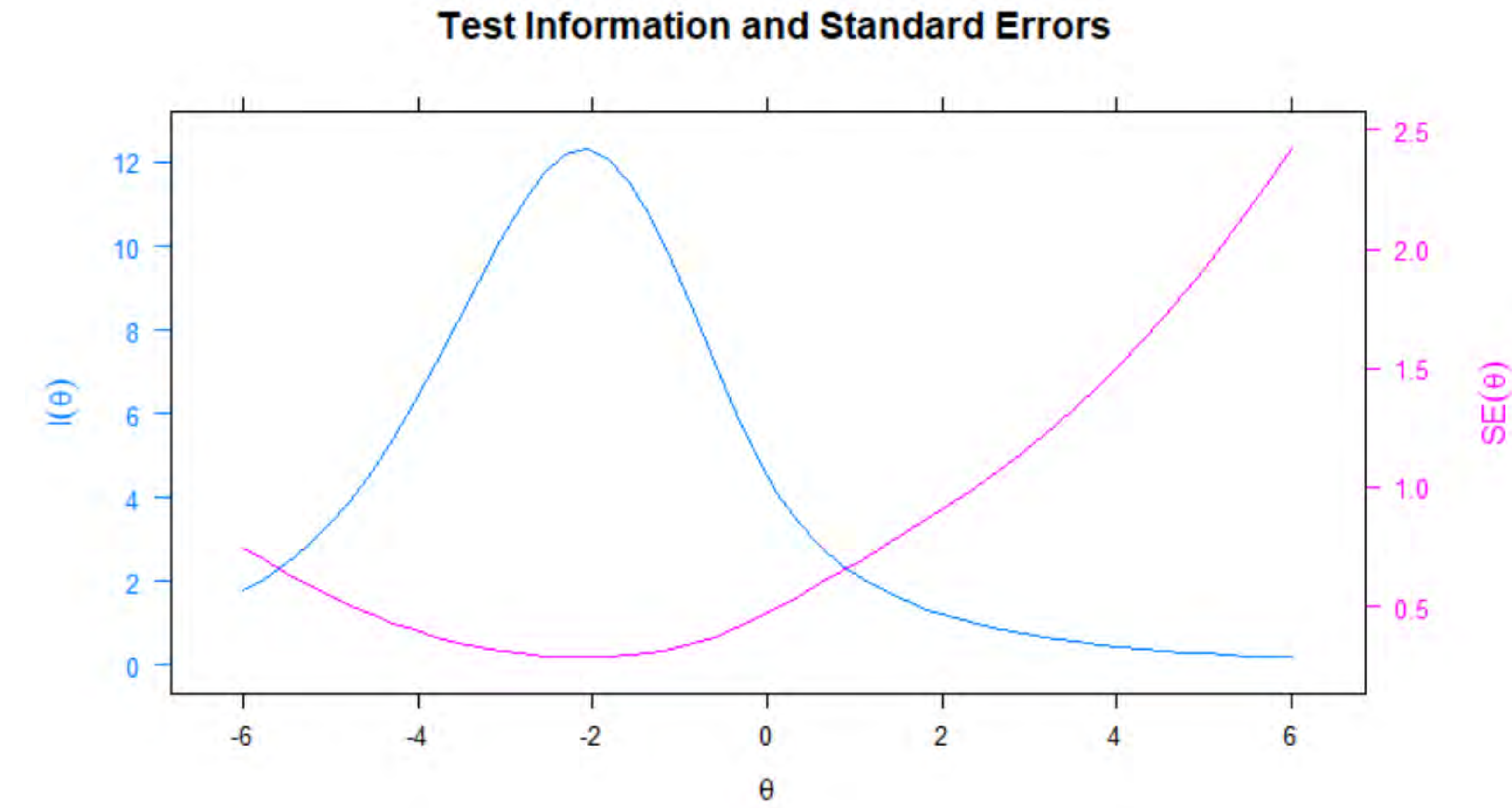
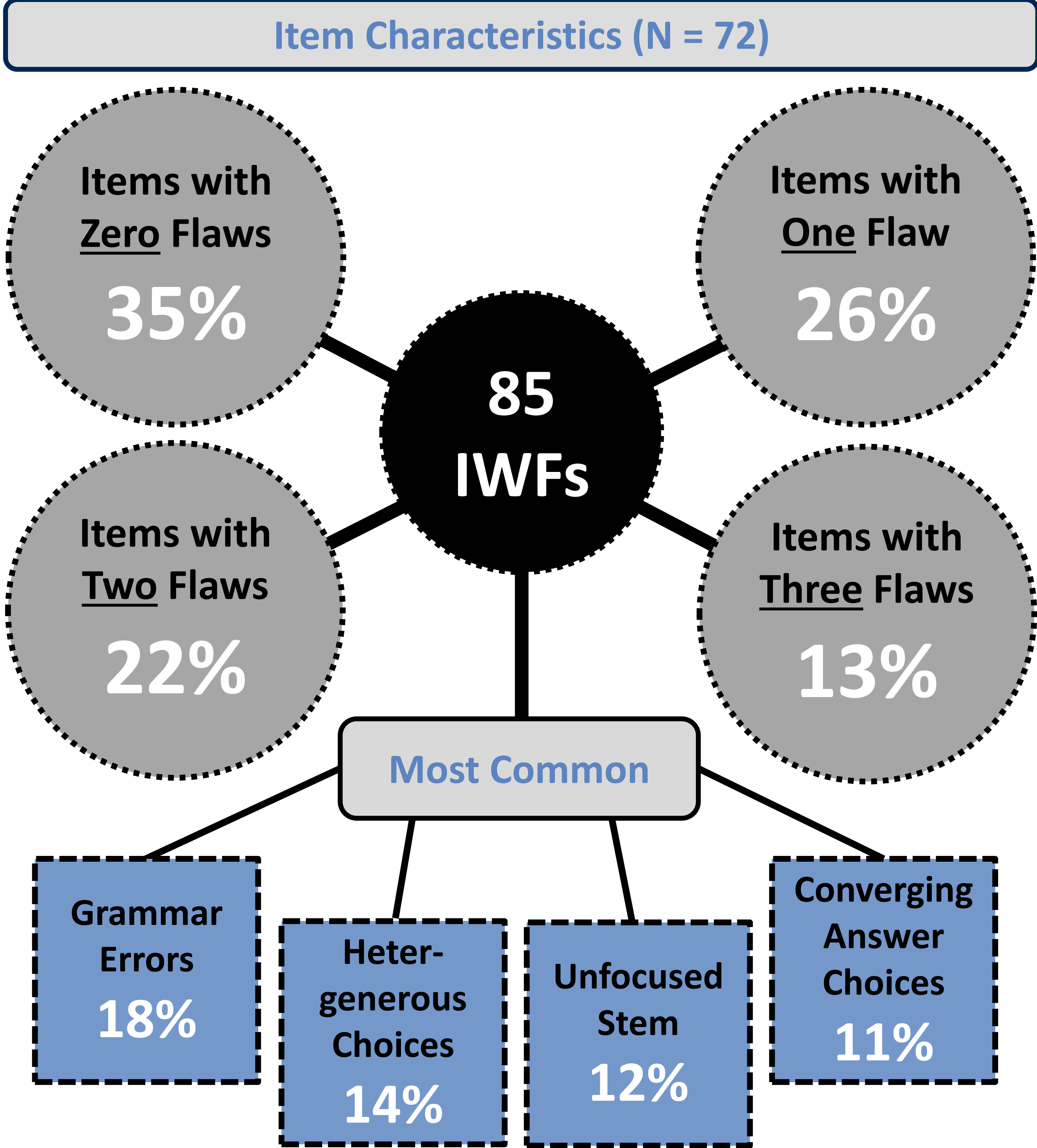


FIGURE 1 Information curve and standard error based on 2P IRT model