Bilingualism: Language and Cognition 22 (5), 2019, 1209–1219 © Cambridge University Press 2018 doi:10.1017/S1366728918000998

Exploring the relationship between multilingualism and tolerance of ambiguity: A survey study from an EFL context*

RINING WEI
Department of English, Xi'an Jiaotong-Liverpool University
YUHANG HU
Department of Linguistics, Georgetown University

(Received: January 23, 2018; final revision received: August 21, 2018; accepted: August 21, 2018)

The relationship between multilingualism and tolerance of ambiguity (TA) has been examined in recent studies (e.g., Dewaele & Li, 2013; van Compernolle, 2016), which focus upon multilinguals with mixed nationalities in non-EFL contexts. Most of these studies regrettably reflect a failure to use effect sizes or provide information on the reliability and validity of the instruments used. The present study explored the relationship between multilingualism and TA by focusing upon 260 English-using multilinguals of one single nationality in an EFL context. Factor analysis revealed a three-factor solution, rather than a four-factor solution of the original TA scale, suggesting a need to re-examine the validity of such instruments when used outside of their native contexts. The results identified multilingualism, number of languages known and gender as important predictors for TA. Given the relative nature of effect-size benchmarks, a topic-specific effect-size benchmark system is proposed to (re-)interpret the present and previous findings.

Keywords: multilingualism, tolerance of ambiguity, factor analysis, validity, reliability, effect size

1. Introduction

Multilingualism has become a powerful fact of life worldwide (Edwards, 2012). In the past decade, research into the effect of multilingualism on individuals' personality has been emerging (cf. Dewaele & van Oudenhoven, 2009; Dewaele, 2012), which is an important complement to the rich ongoing research on the cognitive consequences of multilingualism (cf. Bialystok, Craik & Luk, 2012; Valian, 2015; Bialystok, 2016). Tolerance of ambiguity (TA), a personality variable defined as "the tendency to perceive ambiguous situations as desirable" (Budner, 1962), has been examined vis-à-vis multilingualism in recent years. Most notably, Dewaele and Li (2013) examine the relationship between TA and multilingualism through a large-scale online questionnaire survey: their sample ($N = 2{,}158$) comprised participants from 204 nationalities, with the largest group coming from the USA (n = 478, 22.2% of the total sample) and the 10th largest group from China (n = 41, 1.9%). Most recently, van Compernolle (2016) expands the pioneering research of Dewaele and Li (2013), by introducing a third focal variable, "attitudes toward linguistic variation", and explores the relationships between these focal variables; his sample ($N = 379$) involved respondents from 47 nationalities, with the largest group again coming from the USA (n = 234, 61.7%) and the fifth largest group from the Netherlands (n = 11, 2.9%); although information about participants of Chinese nationality was not given, the percentage of this group was at the most around 2% of his sample. In other words, multilinguals with Chinese nationality (e.g., Chinese users of English)[1] have been under-investigated, as the Chinese population accounts for 20% of the world population whereas only 2% of the samples of the above two studies were Chinese.

Partly motivated by the above gap, this exploratory study aims to examine the relationship between multilingualism and TA, by partially replicating Dewaele and Li's (2013) pioneering work on a group of Chinese users of English in an English as a Foreign Language (EFL) context. This focus brings about three benefits. First, it adds to our understanding of the psychological

* The completion of this paper was made possible thanks to the Research Development Fund of Xi'an Jiaotong-Liverpool University (RDF-16-01-61). Part of this study was presented as a poster at the Harmonious Bilingualism Network (HaBilNet) colloquium (24-25 May, 2018) in Brussels. The authors would like to extend their thanks to the anonymous reviewers and the editors for their constructive comments on an earlier version of this paper. All remaining inadequacies are the authors' responsibility.

Address for correspondence:
Yuhang Hu, Poulton Hall, 1421 37th Street NW, Washington, DC, USA 20057-1051
yh526@georgetown.edu

[1] By "multilinguals" we mean "people with at least partial mastery in a number of languages" (Dewaele & Li, 2013, p. 231). Following Zhao and Campbell (1995), we do not distinguishbetween Chinese "learners" and "users" of English (for the debate on the divide between "learners" and "users", see Kachru, 1992; Yang, 2006). In this study, Chinese users of English are regarded as multilinguals.

profiles of multilinguals in China, an under-investigated context, where the number of English users of Chinese nationality already exceeded 390 million in 2000 (Wei & Su, 2015). Second, examining a sample of multilinguals in the Chinese context represents the first attempt to examine EFL contexts, which provides new data that can complement the extant studies solely from non-EFL contexts (see also Section 2.3). Third, focusing upon a group of multilinguals with one single nationality helps enhance the methodological rigour for this new line of research vis-à-vis the construct validity of TA, as "past studies drew data from a global context and this may not be representative of a single community" (Liu, Wan, Lee & Ng, 2017).

2. Literature review

In our review of empirical studies concerning TA and multilingualism, we argue that several suffer from inadequate use of effect size and/or lack of transparency in instrument reporting. Highlighting major issues concerning effect size use and instrumentation is useful before reviewing studies about TA and multilingualism.

2.1 Methodological rigour

Use of effect size, which is arguably more important than the statistical significance level (i.e., the *p* value) (Ellis, 2010; Larson-Hall, 2010), involves interpreting effect size, which is less straightforward than merely reporting one.

Although Cohen's (1988) benchmarks (e.g., for Cohen's *d*, .20 as small, .50 medium, and .80 large) are widely used for effect size interpretation, these are but general guidelines. As Leech, Barrett, and Morgan (2005, p. 56) note, Cohen's benchmarks, based on the effect sizes usually found in the behavioral sciences, "do not have absolute meaning and are only relative to typical findings in these areas". It is advisable to look for typical values of effect size on the topic of interest, rather than relying on Cohen's rule of thumb (see Wei, Feng & Ma, 2017 for an example of the development of a topic-specific effect-size benchmarks).

Plonsky and Oswald (2014) propose a FIELD-SPECIFIC effect-size benchmark system, and we suggest that TOPIC-SPECIFIC effect-size benchmark systems provide more nuanced guidance in interpreting the effect size in question. On the one hand, a particular field, compared with a particular topic, is broader and more difficult to define; for example, Plonsky and Oswald (2014) do not specify the scope of "L2 research", perhaps because it in itself is "a rather difficult concept to define" (Derrick, 2016, p. 138); it is not clear to what extent the "L2 research" field covers the fields of multilingualism, psycholinguistics, and other L2-related (sub-)fields. On the other hand, Plonsky and Oswald's (2014) benchmarks (e.g., for *r*, .25 as small, .40 medium, and .60 large) in "L2 research" seem to be too high. They are much higher than Cohen's (1988) (e.g., for *r*, .10 as small, .30 medium, and .50 large), probably because the former developed their benchmarks from predominantly experiment-based primary studies, on top of the potential publication bias (Dewaele, 2005) and the "file drawer" problem (Ellis, 2010, p. 69). Experiment-based studies tend to yield higher effect sizes than survey studies, as can be inferred from two sources relevant to TA, which is both a (socio-)psychological and an individual difference variable. First, Richard, Bond and Stokes-Zoota (2003) synthesis, which compiles results from a century of social psychological research covering 25,000 studies of eight million people and featuring a better balance of experiment-based and survey-based studies, finds an average of effect sizes (viz. *r*s) of .21 from this sub-field of psychology, compared with its much higher counterpart (.46) from the field of "L2 research". It is particularly noteworthy that several hundred primary studies concerning the topic of personality yielded an effect size (*r*) average equal to or smaller than .10[2] (Richard et al., 2003), namely Cohen's "small-effect" benchmark. Second, according to our preliminary survey of more recent research (e.g., Cunningham, Douglas & Boag, 2018; Steiger & Reyna, 2017)[3], some predictors in regression models explaining about 1% (equivalent to $r = .10$) of the variance in the dependent variable are regarded as important or "significant predictors". In a recent paper published in *Personality and Social Psychology Bulletin*, Sawyer and Gampa (2018), using 0.1% (roughly equivalent to $r = .03$) as a socially meaningful benchmark to interpret effect sizes, find several statistically significant predictors explaining less than 1% of the variance in the dependent variable. This suggests that 1%-variance-accounted-for variables cannot be simply dismissed as negligible, and their effects need be interpreted in the context of the current understanding of the topic. Hence interpreting effect sizes should be TOPIC-SPECIFIC. The pioneering work on TA by Dewaele and Li (2013) has used effect size to some extent (see Section 2.2 for suggestions). Before we propose a topic-specific effect-size interpretation system (Section 4.3), we will tentatively draw upon Cohen's (1988) system. Reliability and validity measures are

[2] For example, under the topical category of "personality" (Richard et al., 2003: 360), the meta-analytic conclusion that "Intelligent people are popular" (based on 38 primary studies) yielded an average *r* of .10, "Introverts are more vigilant than extroverts" (based on 216 primary studies) an average *r* of .08, and "Sociable, intelligent children are popular with their peers" (based on 176 primary studies) an average *r* of .05.

[3] More similar recent studies can be found in *Personality and Individual Differences* and *Learning and Individual Differences*, two international peer-review journals.

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

classical tools in instrumentation (Chapelle & Duff, 2003; Mahboob, Paltridge, Phakiti, Wagner, Starfield, Burns, Jones & De Costa, 2016). Cronbach's alpha (a measure of internal consistency) is the most frequently used reliability index (Derrick, 2016). A Cronbach's alpha analysis performed on a particular scale assumes that the scale is unidimensional. To check this assumption, exploratory factor analysis is useful. In addition, exploratory factor analysis, which provides information of the construct validity of the instrument (for other types of validity, see Messick, 1995; Brown et al., 2015), should be implemented when examining the factorial structure of an established scale in new cultures (cf. Kim et al., 2011). As Koh, Chang, Fung, and Kee (2007, p. 227) warn, the validity of the scales developed in the West, such as the TA scale by Herman, Stevens, Bird, Mendenhall and Oddou (2010), is "often questionable when they are transported outside of their native land" or context.

2.2 Multilingualism and TA

In multilingualism research, TA has been one of the most frequently examined psychological variables (for others such as extraversion, see Dewaele, 2005; 2012).

An individual with higher TA tends to demonstrate higher ability to (1) take in new information; (2) hold contradictory or incomplete information; and (3) adapt in response to the new information or experience (Ehrman, 1993). TA is highly relevant to second/additional language (L2) learning, which "is often seen as ambiguous" (van Compernolle, 2017, p. 319) because it involves the appropriation of new and/or modified patterns of language and meaning that are usually unfamiliar and complex to the learner. High TA has been considered "essential" to successful L2 learning ever since Rubin's (1975) "good language learner" study (Dörnyei & Ryan, 2015, p. 32), in which it is posited that a "good language learner is . . . comfortable with uncertainty... and willing to try out his guesses" (p. 45).

In the field of multilingualism, TA has been measured with the TA Scale developed by Herman et al. (2010), whereas in related areas (e.g., L2 learning) this personality trait has been assessed frequently with other instruments such as Ely's (1995) Second Language TA Scale (see, e.g., Dewaele & Ip, 2013) and original items by the researchers (e.g., Thompson & Lee, 2013). Herman et al.'s (2010) instrument, developed in what is essentially an English-as-a-native-language or ESL context, is described as "a conceptually clear, internally consistent assessment tool" (p. 60), which is a "refined measure" demonstrating "its improved utility" over Budner's (1962) classic TA inventory (p. 62). Unfortunately, the validity of Herman et al. (2010)'s TA scale has not been fully explored with different contexts and/or populations (see also Section 3.3).

Dewaele and Li's (2013) seminal research on TA and multilingualism draws upon a large group of multilinguals through an Internet-based English-medium questionnaire survey. To assess the respondents' multilingualism and TA, they use a global measure of multilingualism (GMM), viz. "the sum of oral and written knowledge in various languages" (p. 232) and a slightly adapted version of Herman et al.'s (2010) TA instrument. Dewaele and Li (2013) categorise the link between GMM and TA as "weak/small", although not explicitly using Cohen's (1988) benchmark. Specifically, these authors report that (p. 236):

> A one-way ANCOVA with age as a covariate showed that global self-perceived proficiency had a small but significant effect on TA ($F(2,1978) = 6.0, p < .003, \eta^2 = .008$). Age was a significant covariate ($F(1,1978) = 15.1, p < .0001, \eta^2 = .008$). *Post-hoc* pairwise comparisons, with Bonferroni correction, showed that for global self-perceived proficiency the TA scores of the "Low" group were significantly lower ($p < .002$) than those of the "High" group. No significant difference emerged between the Low and Medium group, nor between the Medium and the High groups.

We propose three solutions to overcome shortcomings in reporting and interpreting the ANCOVA results. First, although post-hoc pairwise comparisons are useful, they need to be accompanied by effect size. For example, after mentioning that the difference between the TA scores of the "Low" GMM group and those of the "High" group is STATISTICALLY significant, it is more important to supply an effect size (e.g., *r* as suggested by Field, 2009 [4]). The absence of the term "statistically" can easily result in an erroneous impression that the results are important. Many scholars in psychology (e.g., Carver, 1993) have argued that the term "statistically" must always precede the word "significant"[5].

Second, it is not enough to simply report that no STATISTICALLY "significant difference emerged between the Low and Medium group [sic.]" because even when the result is not statistically significant (or "statistical" in Larson-Hall's terms) the effect size can be large. Therefore, an effect size index should be reported, along

[4] Some books on statistical methods suggest using a measure from the *d* family (e.g. Cohen's *d*). As Larson-Hall (2010, p. 116) points out, in fact either *r* or *d* "can be used". One major benefit of using *r* is that the *r*-family indexes may be easier to understand because their absolute values vary between 0 and 1, unlike the *d*-family indexes (e.g. Cohen's *d*), which may go above 1. This is part of the reason why some meta-analyses (e.g. Richard et al.'s (2003) above-cited synthesis describing one century of social psychological research) employ *r*s.

[5] Some researchers go one step further by suggesting that the word "significant/significance" "should be removed from the statistics vocabulary" (Nassaji, 2012, p. 96). Most recently, Larson-Hall (2016) proposes that the adjective "statistical", rather than the misleading one "significant", be used to denote a "statistically significant" result.

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

with the exact *p* value, regardless of whether the result is statistically significant or not.

Third, the wording "small but significant" potentially diminishes the importance of the finding in Dewaele and Li's (2013) work. Interpreting the effect size (η^2 =.008) as "small" with generic labels (e.g., "medium" and "large") without a reference is a common method[6] to interpret effect size in quantitative studies. But a few interpretative statements, regarding the seemingly "small" η^2 value, would have been useful.

In Cohen's (1988) benchmark system, this effect size (.008) fell below the benchmark (.01) for the so-called "small" effect, suggesting that GMM accounted for .8% of the variance in TA. Although this value was rather "small" according to Cohen's rule of thumb, his generic labels "do not have absolute meaning" (Morgan, Leech, Cloeckner & Barrett, 2004: 90). This effect size, found in Dewaele and Li's (2013) study, may serve as a useful starting point to examine to what extent this value is typical for the effects of sociobiographical variables on TA, so that an effect size interpretation system for this particular topic can be developed.

van Compernolle's (2016) survey, a quasi-replication of Dewaele and Li's (2013) study, also confirms a link between multilingualism (as measured by GMM) and TA, based on a Spearman rho of .19 (p <.0002). Although the correlation coefficient itself (e.g., Spearman rho) represents effect size, "many who use it may not be aware that it is an effect size index" (Ellis, 2010, p. 11). Consequently, the above effect size value (.19) was unfortunately not used to compare with its counterpart (i.e., r = .008) from Dewaele and Li's (2013) article. Furthermore, no information about the reliability[7] or validity of the instrument based upon Herman et al. (2010) was provided.

Liu et al.'s (2017) survey of 132 undergraduate students in Singapore is a recent partial replication of Dewaele and Li's (2013) study. These authors claim that "No significant correlation between global proficiency on TA was found, p = .196". This Singaporean survey is a useful replication study of participants from a single nationality in a non-EFL context. However, the data analysis concentrated on the *p* value, without mention of effect sizes.

Secondly, the Singaporean study fails to report validity and reliability. It misses a valuable opportunity to explore whether the situation "internal consistency of the four dimensions was not sufficiently robust to allow separate use" (Herman et al., 2010, p. 61) occurs, which helps assess the applicability of Herman et al.'s (2010) four-facet TA construct.

The inadequate use of effect size, lack of attention to reliability and validity issues, as well as over-reliance upon participants from non-EFL contexts in previous studies all suggest further (partial) replication studies based on Dewaele and Li's (2013) work.

3. The study

3.1 Research questions

The present study is motivated by the gap concerning TA and multilingualism in under-investigated EFL contexts and the need for stronger methodological rigour in multilingualism research and beyond. It pursues the following questions:

RQ1. What are the underlying factors of the TA scale in the Chinese EFL context?

RQ2. To what extent does the sociobiographical variable, multilingualism (operationalised as GMM), affect TA?

RQ3. To what extent do selected sociobiographical variables other than multilingualism (viz. gender, education, number of languages known, and length of stay abroad) affect TA?

3.2 Participants

A total of 260 Chinese (186 females, 74 males) participated in the present study, ranging from age 18 to 35 (mean = 22.7). Most respondents (n = 195) had or were working towards bachelor degrees, 63 master, and two PhD degrees. Most participants (n = 160) had no experience of living abroad; those with such experience spent an average of 19.63 months (min.: 0.5 month and max.: 14 years; median = 12, mode = 6 months) abroad.

An overwhelming majority (n = 209) of the participants reported to be bilingual, with Chinese as their L1; the others were 41 trilinguals, seven quadrilinguals, two pentalinguals and one sextalingual. The most frequent L2 was English (n = 259) and only one respondent reported Korean as L2. Japanese (n=21) was the most frequent L3, followed by French (n=13), Korean (n=7), Russian (n=3), Spanish (n=2), Germany (n=2) and Portuguese (n=1). In terms of L4, Japanese (n=3) and French (n=3) came first with Korean (n=2) Polish (n=1) and Italian (n=1) following. The pattern for L5 was

[6] The other common methods to interpret effect size include (1) interpretation in relation to one or more relevant studies in the area, and (2) referring to Cohen's (1988) benchmarks.

[7] In another paper by the same author (van Compernolle, 2017), which draws upon the same survey reported in van Compernolle (2016), the reliability indices (measured by Cronbach alpha) are given for the self-developed scales by the author but not for the TA scale adapted from Herman et al. (2010). The author seems to believe that reliability measures are necessary only for self-developed instruments. However, "[i]f using an already established instrument, it is necessary to report the reliability in comparison to the reliability numbers reported in other studies using the same instrument" (Mahboob et al., 2016, p. 50).

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

Japanese (n=1), French (n=1) and Korean (n=1). The only L6 reported was German.

3.3 Instrument

The instrument started with a sociobiographical section comprising conventional questions (e.g., gender, age, education level and length of stay abroad) and a global measure of multilingualism (GMM). The GMM was slightly adapted from the version developed by Dewaele and colleagues (Dewaele & Li, 2013; Dewaele & Stavans, 2014), which has been used in recent studies (e.g., van Compernolle, 2016; 2017). Dewaele and colleagues' original GMM referred to the sum of self-perceived proficiency scores for oral (maximum score 5) and written proficiency (maximum score 5) collected on five-point Likert scales in up to six languages. One major benefit of such a measure is that it is "potentially useful to distinguish sextalinguals with limited knowledge of three languages from trilinguals with advanced knowledge of three languages" (Dewaele & Li, 2014). Dewaele et al.'s GMM thus avoids the lack of clarity inherent to labels such as "bilingual, trilingual", where every language is included, despite the fact that knowledge in some can be very limited. Our only modification of GMM was that the original five-point Likert scale was changed into a nine-point system, as many of our respondents were familiar with the nine-point system used in the IELTS test to elicit more refined linguistic profiles.

Participants' TA was assessed with the TA scale adapted from Herman et al. (2010). The original version was a 12-item questionnaire with five-point Likert scales (1 = "strongly disagree" to 5 = "strongly agree"). It was piloted among 73 Chinese multilinguals. A subsequent reliability analysis of the TA scale revealed that one item dragged the overall Cronbach alpha value down to below .60 (viz. .564). With that item removed, the Cronbach alpha for the pilot test reached .657. Therefore, this item was removed from the final version of the questionnaire; this deleted item in the present study was different from the one deleted in Dewaele and Li's (2013) study (see Appendix 1). Based on feedback from the participants, some minor stylistic adaptations were also made in the final version of the questionnaire.

3.4 Procedures

The anonymous questionnaire was an open-access survey on Wenjuanwang.com, a free China-based survey provider similar to SurveyMonkey.com. Our survey design and questionnaire received ethical clearance from our affiliation. The questionnaire was advertised through several social media. After the pilot-testing, the revised questionnaire was online between January and April, 2016 and attracted 260 valid respondents. Unlike Dewaele and Li's (2013) survey that attracted 2,158 monolinguals and multilinguals, ours did not involve monolinguals because, for all valid respondents, the language of the questionnaire, viz. English, was their foreign language.

Because some respondents left occasional questions blank, the subsample sizes for several variables may vary in the dataset. The dataset was imported into the software package SPSS 22.0 to perform the major statistical procedures.

3.5 Data analysis

RQ1 "What are the underlying factors of the TA scale" was addressed using exploratory factor analysis and reliability analysis. Exploratory factor analysis, rather than its confirmatory counterpart, was chosen because no prior expectations were held regarding the number and nature of underlying factors of the TA scale in the Chinese EFL context.

RQ2, enquiring the extent to which multilingualism affects TA, was attempted with ANOVA and regression, respectively. The ANOVA corresponded to Dewaele and Li's (2013) approach to address a similar question by creating three groups of participants with low, medium and high levels of multilingualism. We followed Plonsky and Oswald's (2017) suggestion that "regression can do everything ANOVA can do, and more", cautioning that "taking a continuous variable and artificially dividing it into two or more groups is a serious mistake". When using ANOVA, we provided a more refined analysis by providing an effect size (r) for each pair-wise comparison (cf. Field, 2009). The absolute value of r ranges between 0 to 1 (the bigger the value, the larger the effect) whereas the squared effect sizes (e.g., an eta squared) give "an underestimated impression of the strength or importance of the effect" (Morgan et al., 2004, p. 90). Hopefully our two statistical procedures are more intelligible.

RQ3, inquiring to what extent selected sociobiographical variables other than multilingualism affect TA, was answered with hierarchical regression, as this statistical procedure helps ascertain the contribution of each predictor variable (Larson-Hall, 2016).

4. Findings and discussion

4.1 The factorial structure of the TA scale

To answer RQ1, the assumptions for factor analysis were first checked. The factorability of the data was checked through the KMO test (.698) and Bartlett's test of sphericity (χ^2 (55) = 437.366, $p < .0005$). These tests and the sample-size-to-variables ratio (23.6) showed that the dataset was appropriate for factor analysis. Principal components analysis was selected for the factor extraction

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

Table 1. *TA by GMM groups (ANOVA).*

		Low Mean (SD)	Medium Mean (SD)	High Mean (SD)	Effect size *r* between each pair (*p*)		
					Low and Medium	Low and High	High and Medium
Item 3	I would like to live in a foreign country for a while.	3.94 (1.181)	4.10 (1.231)	4.51 (.981)	.05 (.476)	.26 (.033)	.12 (.064)
Item 7	If given a choice, I would visit a foreign country rather than vacation at home.	3.84 (1.128)	4.21 (1.193)	4.43 (.850)	.11 (.108)	.29 (.019)	.17 (.193)
Item 8	A good teacher is one who makes you think about/consider your way of looking at things.	4.00 (1.132)	4.16 (1.132)	4.29 (1.073)	.05 (.474)	.12 (.335)	.04 (.545)
TA Core		3.92 (.91)	4.17 (.92)	4.41 (.61)	.09 (.172)	.33 (.015)	.23 (.056)

method, and the direct oblimin rotation was used because it was assumed that the factors would be correlated, which is typical "for naturalistic data, and certainly for any data involving humans" (Field, 2009, p. 644). To extract the most appropriate number of factors, both the Kaiser criterion of using eigenvalues over 1 and the visual inspection of a scree plot were employed. A cut-off point of .40 was adopted for factor loadings (cf. Field, 2009).

Three factors were extracted, accounting for 50.3% of the variance in TA scores (see Appendix 1). The most important finding is that only one factor extracted in this study corresponded to the factorial structure of TA in Herman et al. (2010). This factor, comprising Items 3, 7 and 8, was named "TA core" here, although it had been named "challenging perspectives" by Herman et al. (2010). This name highlights that it may be the very part of TA that could be found across different cultural contexts. No further efforts were made to name the other two extracted factors because of the exploratory nature of this study in EFL contexts. Future studies replicating the TA part of the present study are needed to ascertain to what extent the "TA core" factor is present with different samples of multilinguals in EFL contexts.

A reliability analysis revealed that the Cronbach alpha measure (.30) for the overall TA scale (based upon the 11 items in Appendix 1) was not sufficiently robust to allow the use of the total score to denote TA. However, the internal consistency for the TA core factor (Cronbach alpha = .64) was acceptable, whereas the internal consistencies for the other two factors (.38 and .41 respectively for Factors 1 and 2, see Appendix 1) were not robust enough for separate use. Therefore, in later analysis, TA was denoted by the TA core factor, viz. the average of the scores on Items 3, 7 and 8. The higher the TA score (possible range: 1–5), the higher level of tolerance towards ambiguity that the participant had.

4.2 Multilingualism and TA

Following Dewaele and Li (2013), to answer RQ2 with ANOVA, participants were first divided into three groups (low, medium, high) based on their GMM scores. The participants with scores that were more than 1 standard deviation below the GMM average (M = 29.28, SD = 6.113) were categorised into the "Low" GMM group (n = 31), those with scores that were more than 1 standard deviation above this average into the "High" group (n = 35), and the remaining participants into the "Medium" group (n = 188). A one-way ANOVA test ($F(2, 251) = 2.490$, $p = .085$) revealed that these between-group differences in TA scores were not statistically significant, but the effect size (partial eta squared = .019, $R^2 = .019$), after rounding, reached Cohen's (1988) small benchmark for R^2 (namely .02). To probe further where the differences lay, a series of follow-up t-tests showed that the largest difference ($r = .33$) lay between the Low and High GMM groups, exceeding Cohen's (1988) medium benchmark ($r = .3$), the second largest difference ($r = .23$) existed between the Medium and High GMM groups, and the difference ($r = .09$) between the Medium and Low GMM groups was relatively small, failing to reach Cohen's (1988) small benchmark ($r = .1$) (see Table 1).

Using the ANOVA procedure to answer RQ2 yielded measures readily comparable with those from Dewaele and Li (2013). The patterns we found amongst the three GMM groups, in terms of their mean TA scores, are consistent with those from Dewaele and Li (2013); for

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

example, the largest difference lay between the Low and High GMM groups, which was also reported by Dewaele and Li (2013). But more importantly, we also addressed RQ2, following Plonsky and Oswald's (2017) above-cited suggestion (see Section 3.5), with a simultaneous regression analysis.

Prior to performing the simultaneous regression with the continuous variable GMM as the predictor for TA, we checked that the relevant assumptions (e.g., linearity) had been met. The results show that GMM did not statistically significantly predict TA, F (1, 252) = 3.602, p = .059), but its effect on TA (R^2 = .014, accounting for 1.4% of the TA variance), again, was close to Cohen's (1988) small benchmark.

In a word, the answer to RQ2 is that the effect sizes reflecting the influence of multilingualism on TA explained 1.4% to 1.9% of the TA variance.

Our finding is not in conflict with Dewaele and Li's (2013) finding that GMM "had a small but significant effect on TA". Their p was "< .003" (based on nearly 2,000 participants) and ours ".059" (based on around 250 participants); with a large enough sample size, the p value would always drop below the (arbitrary) conventional level of statistical significance (.05). In the words of authorities on statistics, "surely, God loves the 0.06 nearly as much as the 0.05" (Rosnow & Rosenthal, 1989, p. 1277). Our finding of p = .059 is a case in point to underscore the importance of reporting effect sizes. Although this p value was slightly higher than the conventional statistical significance level adopted (.05), this does not diminish the importance of the result. Readers are advised not to over-emphasise the p value, which is "highly dependent on the sample size" (Mackey & Gass, 2015, p.396); in comparison, however, effect sizes do not fluctuate much with the sample size and hence merit more attention. In connection with Dewaele and Li's (2013) labelling the effect of multilingualism on TA as "small", with more findings concerning the effects of other sociobiographical variables (see RQ3), we will argue below that it would be more useful to develop a topic-specific effect size interpretation system and label the effect of GMM differently.

4.3 Other selected sociobiographical variables and TA

A preliminary analysis was conducted to explore whether the variables of interest in RQ3 could be used as predictors; and, if yes, in what sequence in later hierarchical regression, after the regression assumptions (e.g., normality and homoscedasticity) had been checked? The preliminary analysis confirmed that all the three non-continuous variables and one continuous variable in RQ3 could be used as predictors. Firstly, two independent-samples t-tests demonstrated statistically significant differences of small-to-medium magnitude between males and females (p = .056[8], r = .12), and between bilinguals and "multilinguals"[9] (p = .007, r = .26). Specifically speaking, females (M = 4.239, SD = .864, n = 181) scored higher than males (M = 4.005, SD = .927, n = 73), and "multilinguals" (M = 4.420, SD = .645, n = 50) higher than bilinguals (M = 4.111, SD = .928, n = 204). In other words, both of these non-continuous variables deserved theoretical priority in later regression analysis, where "gender" followed the entry of "number of languages known" because the latter had been shown to be a statistically significant predictor by Dewaele and Li (2013). Secondly, the mean difference between "bachelor degree holders and below" (M = 4.200, SD = .905, n = 190) and those with higher education qualifications (M = 4.089, SD = .832, n = 64) was not statistically significant (p = .386), but the effect size r was .055. The very large p value and the relatively small r led to the tentative hypothesis that "education" would not be a statistically significant predictor for TA; however, this r value, after rounding, still met Cohen's (1988) "small" effect size threshold (.1) and this present study is exploratory in nature because it represents the first attempt in an EFL context to explore the relationship between "education" and TA. Based on these considerations, "education" was also retained for later regression analysis, so as to test the above tentative hypothesis and ascertain its unique contribution to TA. Thirdly, the continuous variable "length of stay abroad" correlated with TA (r = −.057), although this association was not statistically significant (p = .370). Based on similar considerations concerning the variable "education", "length of stay abroad" was also included in later regression analysis to explore its unique contribution to TA.

Table 2 provides the model summary results for the hierarchical regression predicting TA in the model (see Appendix 2 for detailed findings). Each block statistically significantly added to the prediction of the outcome variable (p being .027, .017, .028, and .041, respectively for Blocks 1, 2, 3 and 4). The ΔR^2 column in Table 2 summarises the most important findings: (1) "number of languages known" alone accounted for 1.9% of the variance in TA whereas "gender" accounted for 1.3%, which nearly met the so-called "small" benchmark in Cohen's (1988) system (2%, 13%, and 26% being the small, medium, and large benchmarks); (2) In contrast, the net contributions to the variance in TA by "education" and "length of stay abroad" were .4% and .3%, which were negligible according to Cohen's (1988). Here we have two further examples to illustrate that the p value shall

[8] This is another case in point to underscore the insights in Rosnow and Rosenthal's (1989) above-cited quote. The p value shall never overshadow effect size.

[9] This "non-bilingual" group of multilinguals consists of 41 trilinguals, seven quadrlinguals, two pentalinguals and one sextalingual.

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

Table 2. *Hierarchical Regression Predicting TA: Model Summary.*

	Model 1	Model 2	Model 3	Model 4
	Number of languages known	Gender	Education	Length of stay abroad
R^2	.019	.032	.036	.039
ΔR^2	.019	.013	.004	.003
ΔF	4.946	3.312	.954	.886
p	.027	.017	.028	.041

Note: For Models 2, 3 & 4, the variable underneath 'Model' indicates that it is the newly added variable in this particular model, whereas for Model 1, the variable mentioned is the only predictor in this regression model.

not overshadow effect size; although the *p* values (.028 and .041) fell below the conventional level of statistical significance, it was the effect size value that revealed how important these two variables were in predicting TA.

Our finding in Table 2 that "number of languages known" explained 1.9% of the variance in TA is consistent with Dewaele and Li's (2013) finding that this sociobiographical variable accounted for .09% (p. 236). Another consistent finding is that "education" exerted negligible influence upon TA (i.e., explaining only .4% of the variance), which echoes Dewaele and Li's (2013) result that this biographical variable has no effect on TA.

There are two inconsistencies in our findings when compared with previous research. Firstly, "Stay Abroad", one of the two statically significant predictors of TA in Dewaele and Li (2013), explained 1.4% of the variance. However, its counterpart "length of stay abroad" in our results explained only .3% of the variance and was not statistically significant, suggesting that this sociobiographical variable did not affect TA. This difference could be attributed to the large disparity in the stay abroad experience with these two samples: our sample comprised young multilinguals with much shorter period of stay abroad experience; for example, 17% of the participants had the stay abroad experience of three months or less in this study; in Dewaele and Li's (2013) study, this subgroup (totalling 568 and accounting for 28.557% of the valid 1,989 respondents) was considered as people "who had not lived abroad" (p. 236), suggesting that the majority of their sample had much longer stay abroad experience. Secondly, while Dewaele and Li (2013, p. 235) report that gender exerted "a complete absence of effect on TA", in our study "gender" explained 1.3% of the TA variance, suggesting that gender may be an important predictor for future research. This gender difference merits future efforts to find out whether this difference exists with other samples of Chinese multilinguals or samples of the same nationality in another cultural context.

Given the topic-specific nature of effect size interpretation as discussed in Section 2.1, we propose to develop a benchmark system specifically for interpreting the effects of sociobiographical variables (e.g., GMM) on TA, although we have adopted Cohen's (1988) system thus far. As Dewaele and Li's (2013) study utilised a very large sample and the two identified statistically significant predictors for TA could respectively explain slightly more than 1% of the TA variance, it could have been proposed that $R^2 = .01$ be a typical (or medium) effect size for this line of research. This proposal receives further support from the present study, where the important predictors for TA again respectively accounted for slightly more than 1% of the TA variance; furthermore, an added benefit of using .01 as a benchmark for R^2 is that it corresponds to one commonly used benchmark of its unsquared counterpart ($r = .1$) in Cohen's (1988) traditional system. Based on our findings concerning the respective contribution (viz. below .5% of the TA variance) of two sociobiographical variables to TA, we further propose that $R^2 = .005$ be a small benchmark of effect size. We propose to use $R^2 = .02$, which in Cohen's (1988) system denotes a small effect, as the benchmark for a "large" effect size for interpreting the influence of sociobiographical variables on TA. Theoretically, $R^2 = .09$, which corresponds to another commonly used benchmark of its unsquared counterpart ($r = .3$) in Cohen's (1988) system, can be used a benchmark for a "very large" effect size.

In conclusion, we propose that .005, .01, .02, and .09 be used respectively as the small, typical (medium), large, and very large benchmarks for the effect size R^2 when interpreting the influence of sociobiographical variables on TA. For example, in the present study, the contribution of GMM, number of languages known and gender, to TA respectively exceeded .01; specifically, they accounted for 1.3–1.9% of the TA variance; according to the proposed system, as these effects exceeded the typical benchmark (1% of the variance-accounted-for), these variables can be regarded as important predictors for TA. This benchmark system could potentially be applied to similar lines of survey research focusing upon psychological factors other than TA.

5. Conclusion

The present study has built upon earlier work on multilingualism and TA by focusing upon multilinguals from one particular nationality. The findings from the Chinese EFL context attest to the limitation of the TA scale originally developed by Herman et al. (2010) and support the need for future research on the core of TA in different cultural contexts. Specifically, the TA core identified in this study only contained three items from the original

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

scale, which was claimed to be a "conceptually clear, internally consistent assessment tool" (Herman et al., 2010, p. 60). It would be interesting to see to what extent this TA core can be found with multilingual samples in other replication studies, the value of which is increasingly being recognised (Marsden, Morgan-Short, Thompson & Abugaber, 2018).

In connection with methodological improvements, three suggestions are proposed for future studies. The first is that for sake of higher transparency in instrumentation, researchers should always report the reliability and validity information of their instruments (cf. Derrick, 2016). The second suggestion advocates more adequate use of effect sizes, and a corresponding lower reliance upon the significance level (viz. the *p* value), which has recently been banned by the editors of *Basic and Applied Social Psychology* (Trafimow & Marks, 2015). This journal-wide ban on the use of *p* values represents a natural progression of the long-standing critiques of null hypothesis significance testing (which generates *p*) and a strong call for the employment of more robust statistics (e.g., effect size) in our reporting practices. Concurring with Ellis (2010, p. xiv) who predicts that "If history is anything to go by, statistical reforms adopted in psychology will eventually spread to other social science disciplines", we firmly believe that multilingualism (and the wider field of applied linguistics) will soon be one of these disciplines in Ellis' prediction. The above-proposed effect size benchmarks can be fruitfully employed in studies exploring the effects of sociobiographical variables (e.g., GMM) on TA and possibly on other psychological variables. The third suggestion encourages the use of different measures of multilingualism to examine TA and other psychological variables. To facilitate comparison, this study employed a revised GMM from Dewaele and Li (2013), as a measure of multilingualism; besides GMM, there are other equally useful measures (e.g., Thompson & Khawaja's (2016) operationalisation of multilingualism).

Despite its substantive and methodological contributions, this study has two major limitations. First, it employs the original English-language version of the TA scale. The results were derived from the participants with relatively high proficiency in English. Further research needs to explore the TA of a wider multilingual population, possibly through indigenizing the TA instrument through the translation and back-translation procedure. Second, this study collected data from an online questionnaire, which has its inherent limitations despite its many advantages (Wilson & Dewaele, 2010). It is not clear whether data collected in a more "closed" paper-and-pencil environment, from which the important index of "response rate" can be calculated, could yield different findings. This merits future research efforts.

Appendix 1 *Factor analysis of the TA Scale.*

	Factor		
Factor 1	1	2	3
9. A person who leads an even, regular life in which few surprises or unexpected happenings arise really has a lot to be grateful for.	.703		
10. What we are used to is always preferable to what is unfamiliar.	.670		
5. The sooner we all acquire similar values and ideals the better.	.609		
4. I like to surround myself with things that are familiar to me.	.580		
6. I can be comfortable with nearly all kinds of people.	.538	.415	
11. I like parties where I know most of the people more than ones where all or most of the people are completely strangers.	.446		
Factor 2			
2. I can enjoy being with people whose values are very different from mine.		.768	
1. I avoid situations where people do not share my values.		−.610	
Factor 3			
7. If given a choice, I would visit a foreign country rather than vacation at home.			−.817
3. I would like to live in a foreign country for a while.			−.754
8. A good teacher is one who makes you think about/consider your way of looking at things.	.432		−.662

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

Appendix 2. *Hierarchical Regression Predicting TA: Results.*

	B	*SEB*	*β*	*p*	R^2	ΔR^2	ΔF
Model 1					**.019**	**.019**	**4.946**
Number of languages known	−.309	.139	−.139	.027			
Constant	4.420	.124		.000			
Model 2					**.032**	**.013**	**3.312**
Number of languages known	−.296	.138	−.133	.034			
Gender	.221	.122	.113	.070			
Constant	4.252	.155		.000			
Model 3					**.036**	**.004**	**.954**
Number of languages known	−.303	.139	−.136	.030			
Gender	.220	.122	.112	.072			
Education	−.124	.127	−.061	.330			
Constant	4.166	.178		.000			
Model 4					**.039**	**.003**	**.886**
Number of languages known	−.325	.141	−.146	.022			
Gender	.205	.123	.105	.095			
Education	−.106	.128	−.052	.408			
Length of stay abroad	−.003	.003	−.060	.347			
Constant	4.229	.190		.000			

References

Bialystok, E. (2016). Bilingual education for young children: Review of the effects and consequences. *International Journal of Bilingual Education and Bilingualism*, 1–14. doi: 10.1080/13670050.2016.1203859

Bialystok, E., Craik, F. I., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences, 16*(*4*), 240–250.

Budner, S. (1962). Intolerance of ambiguity as a personality variable. *Journal of Personality, 30*(*1*), 29–50.

Carver, R. P. (1993). The case against statistical significance testing, revisited. *The Journal of Experimental Education, 61*(*4*), 287–292.

Chapelle, C. A., & Duff, P. A. (2003). Some guidelines for conducting quantitative and qualitative research in TESOL. *TESOL Quarterly, 37*(*1*), 157–178.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J.: Lawrence Erlbaum.

Cunningham, M. L., Douglas, H., & Boag, S. (2018). General mental ability moderates the link between confidence and integrity test scores. *Personality and Individual Differences, 123*(*Supplement C*), 94–99.

Derrick, D. J. (2016). Instrument reporting practices in second language research. *TESOL Quarterly, 50*(*1*), 132–153.

Dewaele, J. (2005). Investigating the psychological and emotional dimensions in instructed language learning: Obstacles and possibilities. *The Modern Language Journal, 89*(*3*), 367–380.

Dewaele, J. (2012). Personality: Personality traits as independent and dependent variables. In S. Mercer, S. Ryan & M. Williams (Eds.), *Psychology for language learning* (pp. 42–57). Springer.

Dewaele, J., & Ip, T. S. (2013). The link between foreign language classroom anxiety, second language tolerance of ambiguity and self-rated English proficiency among Chinese learners. *Studies in Second Language Learning and Teaching, 3*(*1*), 47–66.

Dewaele, J., & Li, W. (2013). Is multilingualism linked to a higher tolerance of ambiguity? *Bilingualism: Language and Cognition, 16*(*1*), 231–240.

Dewaele, J., & Li, W. (2014). Attitudes towards code-switching among adult mono- and multilingual language users. *Journal of Multilingual and Multicultural Development, 335*(*3*), 235–251.

Dewaele, J., & Stavans, A. (2014). The effect of immigration, acculturation and multicompetence on personality profiles of Israeli multilinguals. *International Journal of Bilingualism, 18*(*3*), 203–221.

Dewaele, J., & Van Oudenhoven, J. P. (2009). The effect of multilingualism/multiculturalism on personality: No gain without pain for third culture kids? *International Journal of Multilingualism, 6*(*4*), 443–459.

Dörnyei, Z., & Ryan, S. (2015). *The psychology of the language learner revisited*. London: Routledge.

Edwards, J. (2012). *Multilingualism: Understanding linguistic diversity*. London: Bloomsbury Publishing.

Ehrman, M. E. (1993). Ego boundaries revisited: Toward a model of personality and learning. In J. E. Alatis (Ed.), *Georgetown university round table on languages and linguistics 1993* (pp. 330–362). Washington: Georgetown University Press.

https://doi.org/10.1017/S1366728918000998 Published online by Cambridge University Press

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge: Cambridge University Press.

Ely, C. M. (1995). Tolerance of ambiguity and the teaching of ESL. In M. Reid Joy (Ed.), *Learning styles in the ESL/EFL classroom* (pp. 87–95). Boston: Heinle & Heinle.

Field, A. P. (2009). *Discovering statistics using SPSS: (and sex and drugs and rock 'n' roll)* (3rd ed.). Los Angeles: SAGE.

Herman, J. L., Stevens, M. J., Bird, A., Mendenhall, M., & Oddou, G. (2010). The tolerance for ambiguity scale: Towards a more refined measure for international management research. *International Journal of Intercultural Relations, 34*(*1*), 58–65.

Kachru, B. B. (1992). Teaching world Englishes. In B. B. Kachru (Ed.), *The other tongue, English across cultures* (pp. 355–365). Urbana: University of Illinois Press.

Koh, J. B., Chang, W. C., Fung, D. S., & Kee, C. H. (2007). Conceptualization and manifestation of depression in an Asian context: Formal construction and validation of a children's depression scale in Singapore. *Culture, Medicine and Psychiatry, 31*(*2*), 225–249.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS*. New York: Routledge.

Larson-Hall, J. (2016). *A guide to doing statistics in second language research using SPSS and R*. New York: Routledge.

Leech, N. L., Barrett, K. C., & Morgan, G. A. (2005). *SPSS for intermediate statistics: Use and interpretation*. Mahwah: Lawrence Erlbaum Associates.

Liu, T., Xuan, J., Lee, Y., & Bee Chin. (2017). What types of multilinguals are more tolerant of ambiguity? the role of multilingualism and attitudes towards linguistic variation. *Paper Presented at the11th International Symposium on Bilingualism (ISB11)*, Limerick, Ireland.

Mackey, A., & Gass, S. M. (2015). *Second language research: Methodology and design* Routledge.

Mahboob, A., Paltridge, B., Phakiti, A., Wagner, E., Starfield, S., Burns, A., Jones, R.H., & De Costa, P. I. (2016). TESOL quarterly research guidelines. *TESOL Quarterly, 50*(*1*), 42–65.

Marsden, E., Morgan-Short, K., Thompson, S., & Abugaber, D. (2018). Replication in second language research: Narrative and systematic reviews and recommendations for the field. *Language Learning, 68*(*2*), 321–391.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(*9*), 741.

Morgan, G. A., Leech, N. L., Cloeckner, G. W., & Barrett, K. C. (2004). *SPSS for introductory statistics: Use and interpretation* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum.

Nassaji, H. (2012). Statistical significance tests and result generalizability: Issues, misconceptions, and a case for replication. In G. Porte (Ed.), *Replication research in applied linguistics* (1st ed., pp. 92–115). New York: Cambridge university press.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition, 39*(*3*), 579–592.

Plonsky, L., & Oswald, L. F. (2014). How big is "big"? interpreting effect sizes in L2 research. *Language Learning, 64*(*4*), 878–912.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(*4*), 331–363.

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*(*10*), 1276–1284.

Rubin, J. (1975). What the" good language learner" can teach us. *TESOL Quarterly, 9*(*1*), 41–51.

Sawyer, J., & Gampa, A. (2018). Implicit and explicit racial attitudes changed during black lives matter. *Personality and Social Psychology Bulletin, 44*(*7*), 1039–1059.

Steiger, R. L., & Reyna, C. (2017). Trait contempt, anger, disgust, and moral foundation values. *Personality and Individual Differences, 113*(*Supplement C*), 125–135.

Thompson, A. S., & Lee, J. (2013). Anxiety and EFL: Does multilingualism matter? *International Journal of Bilingual Education and Bilingualism,16*(*6*), 730–749.

Thompson, A. S., & Khawaja, A. J. (2016). Foreign language anxiety in Turkey: The role of multilingualism. *Journal of Multilingual and Multicultural Development,37*(*2*), 115–130.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology, 37*(*1*), 1–2.

Valian, V. (2015). Bilingualism and cognition. *Bilingualism: Language and Cognition*, *18*(*1*), 3–24.

van Compernolle, R. A. (2016). Are multilingualism, tolerance of ambiguity, and attitudes toward linguistic variation related? *International Journal of Multilingualism,13*(*1*), 61–73.

van Compernolle, R. A. (2017). Preferences for (in)formal language: Correlations with attitudes toward linguistic variation, multilingualism, tolerance of ambiguity, and residence abroad. *International Journal of Multilingualism, 14*(*4*), 317–331.

Wei, R., Feng, J., & Ma, Q. (2017). College students' perspectives on English-medium instruction and their English learning motivational intensity. In J. Zhao, & Q. Dixon (Eds.), *English-medium instruction in Chinese universities: Perspectives, discourse and evaluation* (pp. 45–58). London: Routledge.

Wei, R., & Su, J. (2015). Surveying the English language across China. *World Englishes, 34*(*2*), 175–189.

Wilson, R., & Dewaele, J. (2010). The use of web questionnaires in second language acquisition and bilingualism research. *Second Language Research, 26*(*1*), 103–123.

Yang, J. (2006). Learners and users of English in China. *English Today, 22*(*2*), 3–10.

Zhao, Y., & Campbell, K. P. (1995). English in China. *World Englishes, 14*(*3*), 377–390.